Методические рекомендации пользователю системы мониторинга состояния межнациональных отношений по данным социальных сетей

С.Н.Кольцов, О.Ю.Кольцова
ЛИНИС ВШЭ СПб
Ноябрь 2017

Общие рекомендации по программным средствам.

Система мониторинга межнациональных отношений по данным социальных сетей реализована в двух видах: 1. Оффлайновая система ‘TopicMiner’, реализованна в виде 64-битного программного средства для Windows 10. 2. Онлайновая система ‘Web TopicMiner’, реализована в виде сайта, который может работать под операционными систеами Windows и Unix. Оффлайновая система не требует специальной установки. Её достаточно скопировать на компьютер. Однако при использовании рекомендуется установить дополнительное бесплатное обеспечение Microsoft Visual C++ 2015, Redistributable X64. Все подробности по установке программных средств приведены в руководстве пользователя. Рекомендуется перед установкой ПО ‘TopicMiner’ ознакомится с соответствующими главами из руководства пользователя. Онлайновая система ‘Web TopicMiner’ требует процесса инсталляции как для Windows, так и для Unix. Поэтому процедуру инсталляции и настройки доступа к ‘Web TopicMiner’ необходимо проводить силами сотрудников, обладающих соответствующими профессиональными навыками. Однако, сама эксплуатация онлайновой системы мониторинга не требует специализированных знаний и может осуществляется сотрудниками, умеющими работать с web браузерами.

Задачи, решаемые с помощью системы мониторинга.

Система мониторинга, прежде всего, предназначена для извлечения из миллионов документов, при помощи специализированных алгоритмов, нужных тем и связанных с ними документов. Исходя из этого, с помощью данной информационной системы можно решать следующий список задач: 1. Тематическое картирование как русскоязычных, так и англоязычных социальных сетей. 2. Определение медийного контента телевизионных каналов и газет, представленных в виде многомиллионных текстовых коллекций. 3. Выделение текстовых коллекций на заданную оригинальную тему. 4. Проведение тонального анализа выделенных тем. 5. Анализ географического распределения тематической представленности на карте Российской Федерации. 6. Применение всех выше описанных функций для анализа этнической напряженности в социальных сетях.

Пользователям важно понимать, чего данная система не делает. Во-первых, она не собирает данные, представляя собой аналитический модуль. Данные следует либо собирать самостоятельно, либо через существующие сервисы агрегации текстов социальных сетей и СМИ. Во-вторых, система ищет в коллекции текстов естественно существующие в них темы, но не проверяет гипотезы пользователей о наличии в коллекциях предзаданных тем. Это связано с тем, что система создана для задач выявления латентных тем, о которых пользователь может и не подозревать. В-третьих, система может помочь в определении тональной окрашенности обнаруженной темы в целом, но не отдельных текстов. Это опять же вытекает из задачи системы, связанной с мониторингом трендов в целом. В-четвертых, система напрямую не отвечает на вопрос, хорошо или плохо относятся пользователи соцсетей к той или иной этнической группе. Вместо этого она может: (а) найти связанную с данной этнической группой тему (темы), если этническая группа вообще обсуждается; (б) помочь пользователю быстро определить содержание этой темы, то есть контекст обсуждения данной этнической группы; (в) помочь пользователю быстро определить общую тональность темы (при этом негативная тональность будет признаком проблематичности темы); (г) отследить, растет или

уменьшается обсуждение этой темы со временем, и в каком регионе оно наиболее активно. Наконец, в-пятых, поскольку система работает по принципу «мешка слов», то есть все алгоритмы основаны на анализе встречаемости слов в текстах, она несет в себе все ограничения данного подхода.

Рекомендуемая последовательность действий и общая логика мониторинга.

Мониторинг межэтнических отношений с использованием нашей системы предполагает следующую логику и соответствующую ей последовательность действий.

Шаг 1. Сбор данных из социальных сетей. Возможен как при помощи ПО 'VkMiner', так и коммерческих агрегаторов: 'IqBuzz', Wobot, Brand Analytics и др.

Шаг 2. Парсинг и унификация геогрифческих названий в коллекции. Шаг нужен, т.к. мультимодальное тематическое моделирование с учетом метаданных (дата, геотег) и визуализация результатов тематического моделирования на карте РФ требует унификации метаданных. Поэтому к системе предлагается отдельный скрипт для унификации, а также список регионов и городов.

Шаг 3. Препроцессинг текстовых данных (модуль в оффлайновой системе). На данном шаге необходимо производить лемматизацию, разработку списка стоп-слов, и, если необходимо, удалять стоп-слова из лемматизированной коллекции. Для упрощения работы с информационной системой реализован полный цикл препроцесиинга. Результат препроцессинга хранится в формате tmlda, который поддерживается как оффлайновой системой, так и онлайновой системой мониторинга. Кроме того, в модуль препроцесинга и ремонта данных встроен алгоритм унификации дат.

Шаг 4. Фильтрация текстов, нерелевантных теме этничности. Осуществляется с помощью функции классификации по параметру «релевантный / релевантный» в онлайновой системе. В эту опцию встроена своя функция лемматизации, так что ее можно запускать перед шагом 3.

Шаг 5. Выявление контекстов упоминания этнических групп с помощью тематического моделирования. Выбор моделей и их параметров зависит от задач и опиан ниже. За построением моделей следует сортировка слов и текстов внутри тем по их вероятности и ручное присвоение названий темам на основе чтения топ-слов и топ-текстов в каждой теме. Выявление этнорелевантных тем может быть ускорено за счет функции подсказки, подкрашивающей соответствующие темы в определенные цвета (на основе словаря).

Шаг 6. Определение распределения тем во времени и пространстве. Можно создать график изменения активности обсуждения выбранной темы во времени и визуализировать веса темы за определенный период на карте РФ, по субъектам федерации.

Шаг 7. Анализ тональности. Система не выявляет напрямую отношения к конкретной группе, но имеет несколько опций, позволяющих оценить их косвенно. А. Функция тональности темы исходя из ее топ-слов. Б. Функция классификации текстов на содержащие и не содержащие общий негативный сентимент. В. Функция классификации текстов на содержащие и не содержащие упоминания этнического конфликта. Г. Возможность выгрузить тексты во внешнее ПО SentiStrength и провести сентимент-анализ на основе словарного подхода с нашим словарем.

Шаг 8. Выгрузка результатов во внешние файлы.

Рекомендации по сбору данных.

Решение описанных задач системой зависит от нескольких важных параметров, связанных со сбором данных. Прежде всего, необходимо обеспечивать полноту текстовых коллекций. Можно было бы собирать все тексты из социальных сетей, однако опыт показывает, что такие коллекции настолько зашумлены и доля в них этнорелевантных текстов настолько мала, что они не поддаются анализу никакими алгоритмами. К шуму относятся, в первую очередь, короткие тексты; тексты, созданные из тредов комментариев

к постам; тексты, наполненные списками слов для СЕО-оптимизации и другие. Кроме того, может существовать большое количество дубликатов, которые для некоторых задач нужно отсеивать. Рекомендации по очистке коллекции от дубликатов приводятся ниже.

Таким образом, для данной информационной системы, коллекции необходимо формировать специальным образом. Сбор данных для мониторинговой системы можно реализовать двумя способами:

1. **Сбор данных при помощи программного средства VKMiner**. Данное программное средство входит в состав мониторинговой системы в виде дополнительного программного средства. Подробное описание VKMiner приведено в руководстве пользователя. Достоинством данного ПО является отсутствие платы за его использование и возможность таргетированного сбора данных, то есть сбора данных с конкретных страниц пользователей или групп в социальной сети. Недостатком данного ПО является сложность сбора действительно больших данных и отсутствие доступа к другим социальным сетям, таким как Facebook, Twitter, Одноклассники и другие.

2. **Сбор данных при помощи коммерческих агрегаторов**, таких как Iqbuzz или YouScan. Достоинством коммерческих агрегаторов является возможность получения больших данных с метаданными из разных социальных сетей, а недостатком является достаточно высокая стоимость при регулярном сборе данных для системы мониторинга. Кроме того, как показали наши исследования, ≅80% этнорелевантных текстов, собранных с помощью агрегатора Iqbuzz, составляют данные из социальной сети Вконтакте. Также, с целью получения существенной доли этнически окрашенных документов в коллекции рекомендуется использовать список этнических групп при закачивании данных с помощью коммерческих агрегаторов. Разработанный список этнонимов для мониторинга пост-советского пространства приложен к информационной системе.

Рекомендации по препроцессингу данных.

Результаты мониторинга зависят от того, как данные были обработаны, прежде чем они были поданы в систему анализа. Так, в целом на тематическое моделирование отрицательно влияет наличие дубликатов. Казалось бы, если пользователи социальных сетей часто перепостят какой-то текст, его множественное появление должно быть учтено и должно привести к образованию крупной темы. Однако на практике происходит следующее: тема вокруг перепоста действительно образуется, но все остальные похожие тексты от нее откалываются. В таких случаях можно рекомендовать: (а) провести тематическое моделирование без удаления дубликатов и с удалением и сравнить; (б) провести тематическое моделирование после удаления дубликатов, а информацию о количестве перепостов у сообщения использовать отдельно. Если сообщение перепостили значимое количество раз, и оно конфликтно, это уже сигнал для аналитика даже без тематического моделирования.

Для удаления дубликатов рекомендуется использовать технологию ‘шинглов’. Данная технология реализована в виде интерфейсного программного средства и является дополнительным программным средством в нашей системе. Подробности применения данного ПО приведено в руководстве пользователя. Удаление дубликатов позволяет уменьшить исходную коллекцию практически в несколько раз, что существенно для мониторинга, так как резко сокращает время препроцессинга и проведения тематического моделирования. Так, в ходе разработки системы в IQBuzz - компании, собирающей данные из всех русскоязычных социальных сетей – были собраны всех тексты, в которых упоминалась хотя бы одна этническая группа пост-советского пространства за два года. Она составила около 5,5 млн. текстов. После удаления всех дубликатов осталось менее 2,5 млн. текстов.

Другим существенным фактором, влияющим на результат работы информационной системы мониторинга, является качество препроцессиннга данных. В основе препроцессинга лежит процедура лемматизации и удаление стоп-слов. Лемматизация –

это приведение всех слов к начальной форме для правильного подсчета их частот. Процедура лемматизации реализована при помощи программных средств разработанных компанией Яндекс и поэтому не требует рекомендаций. Однако необходимо тщательно подходить к процессу удаления стоп-слов. Стоп-словами называют слова, мешающие работе алгоритма. Рекомендуется, прежде всего, удалять наиболее частотные слова и слова, которые встречаются менее 3-5 раз в коллекции. Также рекомендуется использовать уже разработанный список стоп-слов, которые входит в состав информационной системы. Данный список составлен на основе русскоязычной коллекции документов, собранных из социальных сетей. Кроме того, рекомендуется не удалять слова, если они относятся к темам, которые нужно мониторить, например этнонимы, название населенных пунктов, регионов и иные геотеги, даже если они встречаются слишком часто или слишком редко. У пользователей также могут быть свои гипотезы о том, как те или иные слова связаны с контекстом упоминания этнонимов или с их тональностью. Например, личные местоимения (я, мы, вы) обычно удаляются, но в некоторых случаях они могу играть важную роль (например, при противопоставлении типа «мы-они» в этническом конфликте), и тогда их удалять не следует. Необдуманное удаление стоп-слов может привести к исчезновению нужных тем.

Рекомендации по повышению доли этнорелевантных текстов в коллекции.

При разработке системы мониторинга было обнаружено, что доля текстов, имеющих отношение к этничности, в общем потоке пользовательских текстов ничтожно мала. Большинство текстов посвящено повседневной тематике. Как следствие, тематическое моделирование не работает: оно просто не в состоянии вычленить темы, встречающиеся в ничтожной доле текстов. Единственным выходом из ситуации является повышение доли релевантных текстов в коллекции. Это можно сделать двумя способами: поиском текстов по ключевым словам и автоматической классификацией текстов на основе машинного обучения.

Поиск по ключевым словам. Первый недостаток этого метода в том, что задавая ключевые слова, пользователь может внести свои субъективные представления релевантности текстов, которые могут быть ошибочными. Например, может оказаться, что слово «национализм» приводит нас только к текстам, его изобличающим, уводя поиск от текстов, разжигающих межнациональную рознь. Поэтому мы составили список ключевых слов и словосочетаний, содержащих только названия этнических групп и этнических персонажей пост-советского пространства (напр., еврей, жид, кавказец, хохлушка, «русский народ», «татарская девочка»). Этот список прилагается к нашей системе. При его использовании будет неизбежен второй недостаток этого подхода: в коллекцию не попадут тексты, обсуждающие проблемы межэтнических отношений без упоминания конкретных этических групп, например, в терминах «понаехали тут всякие». Однако плюс этого подхода в том, что подавляющее большинство – около 80% текстов, попадающих в такую коллекцию, - действительно имеют отношение к этничности. То есть этот метод имеет высокую точность, не неизвестную полноту.

Для улучшения полноты можно использовать метод автоматической классификации с помощью классификатора, специально обученного находить этнорелевантные тексты. Он встроен в систему мониторинга в виде функции, позволяющей запустить внешнюю программу. Однако следует отметить, что этот модуль еще требует доработки. Если полнота предпочтительна – а это, как правило, так и есть для задач мониторинга этнических конфликтов – рекомендуется включать в коллекцию как все тексты, отобранные по нашему словарю, так и все, отобранные с помощью классификатора.

Рекомендации по обработке метаданных.

Для тематического моделирования рекомендуется использование метаданных, то есть дополнительных характеристик текстов, таких как дата, место, характеристики автора. Это возможно двумя способами. Во-первых, их можно использовать после завершения тематического моделирования. Если суммировать вероятности темы по всем текстам какого-то типа – например, определенной даты, региона или, скажем, пола автора, можно получить выраженность темы в данной группе текстов, то есть то, насколько активно она обсуждалась в данное время, в данном месте, пользователями данного пола. В нашей системе мониторинга автоматически можно рассчитать веса тем по субъектам Российской Федерации и визуализировать каждую выбранную тему по отдельности на карте Российской Федерации. Во-вторых, метаданные можно использовать непосредственно в мультимодальном тематическом моделировании и получить распределение геотегов по темам. Такой подход сделает выраженность тем по регионам более выпуклой, однако он «накажет» темы, которые равномерно обсуждаются везде, даже если они обсуждаются много. То есть, общероссийские темы станут несколько менее выраженными. Учет исходных метаданных реализован в виде дополнительного файла в формате ‘csv’. Особенности формирования файла метаданных описаны в руководстве пользователя.

Необходимо отметить, что в сырых исходных метаданных много мусора. Пользователей социальных сетей часто либо пропускают указание своего города или региона, либо указывают свое местоположение с произвольной форме, например, ‘планета_земля_моя_могучая_россия_набережные_челны’. Поэтому, для того чтобы в дальнейшем иметь возможность применять геотеги в тематическом моделировании, рекомендуется провести процедуру унификации городов и регионов в метаданных документов. Для этой цели можно использовать программное обеспечение ‘gtp.exe’, которое входит в состав нашей системы как дополнительное. Данное ПО не требует установки. Целью работы этого программного средства является сравнение метаданных с БД из Вконтакте. В результате получаются унифицированные данные, которые можно использовать в системе мониторинга.

Рекомендации по проведению тематического моделирования.

Рекомендации по проведению тематического моделирования делятся на следующие части: 1. Выбор модели. 2. Определение числа тем. 3. Интерпретация результатов тематического моделирования. 4. Рекомендации по проведению сентимент-анализа.

Выбор модели. В нашей системе реализовано несколько моделей; выбор модели определяется, прежде всего, поставленной задачей. Если требуется определить общую тематическую структуру заданной коллекции, без выделения специфичных тем, следует использовать модели LDA Gibbs sampling или pLSA в реализации BigARTM. Подробности применения этих моделей рассмотрены в руководстве пользователя ПО ‘TopicMiner’. Если пользователь уверен, что в интересующих его темах должны обязательно быть определенные слова (и он точно знает, какие), рекомендуется использовать так называемые алгоритмы с частичным обучением: либо ISLDA Gibbs sampling или MultiModels в реализации BigARTM. Например, пользователя могут интересовать темы, для которых характерна высокая вероятность употребления каокй-либо торговой марки или, как в нашей случае, определенного этнонима. В этом случае этничность задается в виде совокупности специальных слов которые распределяются по фиксированным темам. ISLDA требует закрепления слова или нескольких слов за каждой темой в отдельности; BigARTM может закрепить сразу весь список слов за диапазоном тем (например, сразу да двадцатью или ста темами). В обоих случаях, в ходе моделирования, в зафиксированных темах, в которых, верхушка задана этнонимами, также появляются слова, заранее не известные пользователю, но имеющие отношение (с высокой вероятностью) к заданным

этнонимам. Они и будут признаками, определяющими контекст употребления заданного слова. К нашей системе прилагается список этнических групп пост-советского пространства. Однако если пользователю требуется изучить, как освещаются в блогах народы Европы или другие сущности, ему придется составить свой список слов. Если у пользователя нет никаких гипотез о том, какие слова должны обязательно быть высоко вероятными по интересующим его темам, модели с частичным обучением использовать не стоит.

Если требуется получить распределение этнической напряженности с учетом метаданных, следует использовать только MultiModels, так как только данная модель учитывает заданные поля метаданных (такие как время, город, регион) непосредственно в тематической модели. Остальные модели позволяют визуализировать результаты тематического моделирования с учетом геотега, но сами геотеги не включены в тематическое моделирование.

Определение числа тем. Проблема выбора числа тем в тематическом моделировании пока не решена на теоретическом уровне. Однако есть ряд практических рекомендаций по подбору оптимального числа тем. Во-первых, исходя из опыта, при поиске этнорелевантных тем рекомендуется использовать число тем порядка 200-300 на коллекции от 100 000 документов. Во-вторых, не рекомендуется использовать число тем более 400, так как дальнейшее увеличение числа тем приводит к тому, что распределения слов и документов по темам стремится к равномерному, происходит сильная флуктуация слов с маленькой вероятностью, и на поверхность в темах всплывают мусорные слова, не отражающие значимые и интересные темы. В-третьих, число тем зависит от требуемого уровня обобщения. Так, 50 тем на коллекции в 100 000 газетных статей будут очень общими, такими как: «спорт», «культура», «здравоохранение», «выборы». Кроме того, существует еще один фактор, который влияет на выбор числа тем: это так называемая семантическая стабильность. Под семантической стабильностью понимается ситуация когда тема воспроизводится при множественном запуске программы тематического моделирования, на одной и той же коллекции и с одними и теме же параметрами. Рекомендуется производить запуск ‘TopicMiner’ три раза, затем использовать опцию расчета меры Кулбака–Лейблера для определения стабильных тем. Подробности определения семантически стабильных тем описаны в руководстве пользователя. Проблема семантической стабильности существует потому, что в основе тематического моделирования лежать процессы генерации случайного чисел. Это приводит к флуктуациям вероятностей слов по темам при разных запусках.

Интерпретация результатов тематического моделирования. Результатами тематического моделирования является матрицы распределения слов и документов по темам, а также, в случае использования мультимодальных схем расчета, матрицы распределения метаданных по темам. То есть, пользователь, во-первых, видит таблицу выраженности разных тем в текстах, но по ней он пока не может определить, о чем эти темы. Во-вторых, он видит таблицу типичности слов в темах: просмотрев самые типичные слова, он увидит, о чем тема. В-третьих, он видит данные выраженности тем в группах текстах – например, региональных – в зависимости от того, какие метаданные были поданы в систему. Прежде чем анализировать результаты расчета, рекомендуется провести сортировку матриц. Таким образом, лучше всего анализировать наиболее вероятностные слова, документы, геотеги по темам. Рекомендуется анализировать первые 20-50 слов, отсортированных по вероятности, так как вероятность слов в темах падает достаточно быстро. При анализе наиболее вероятностных слов следует обратить внимание на интерпретируемость тем, так как тематическое моделирование может выделять темы, которые не имеют смысла, например, в теме могут фигурировать одни имена. В случае плохой очистки, в состав топ слов могут входить и предлоги, и междометия. Если вероятность даже топовых слов в теме низка, значит эта тема слабо выражена.

В силу того, что люди могут распознавать тему в топовых словах по-разному, рекомендуется оценивать тему тремя разными пользователями. Согласованность ("**Re**liability **Cal**culator for **O**rdinal, **I**nterval, and **R**atio data") между кодировщиками по трем оценкам для каждого слова можно рассчитать при помощи онлайн калькулятора (http://dfreelon.org/utils/recalfront/recal-oir/), либо с помощью других калькуляторов.

Чтение наиболее вероятных текстов в теме может быть использовано как для лучшего понимания смысла темы, так и для более детального изучения самих текстов. Бывает так, что тема, не понятная на основе топ-слов, может стать гораздо понятнее после чтения текстов. К сожалению, бывает и наоборот (особенно для тем, чьи топ-слова имеют низкие вероятности).

Если все темы получились неинтерпретируемыми, это может быть связано со следующими причинами: 1. Тексты плохо очищены. 2. Тексты плохо лемматизированы. 3. Почти все тексты слишком короткие (1-2 предложения и менее). 4. Тексты почти не отличаются друг от друга по составу слов. 5. Количество тем слишком велико. 6. Количество итераций слишком мало. 6. Стандартные значения параметров альфа и бета не подходят для данной коллекции. Проблемы 1, 2, 5 многие пользователи могут решить самостоятельно; с проблемой 6 лучше обратиться к специалисту по тематическому моделированию, например, к разработчикам данной системы. Проблемы 3 и 4 означают, что тематическое моделирование – не подходящий метод для имеющихся коллекций.

Рекомендации по проведению сентимент-анализа.

Полный сентимент-анализ представляет собой определение тональности как слов, так и текстов в каждой теме. Это можно делать на основе двух подходов: машинного обучения и словарного подхода. При исследовании социальных явлений, таких как освещение этничности, по ряду причин словарный подход предпочтителен. Поэтому в нашей системе реализован словарный подход, где качестве словаря предлагается наш словарь, созданный специально для работы с пользовательскими текстами социально-политической тематикив рамках проекта 'Разработка общедоступной базы данных и краудсорсингового веб-ресурса для создания инструментов сентимент-анализа № 14-04-12031'. Внутрь системы интегрирована только функция определения общей тональности темы по словам. Определить тональность текстов с помощью нашего словаря можно с помощью внешнего ПО SentiStrength, для которого он разработан (Правила использования 'Sentistrenth' приведено в руководстве пользователя).

Наша система размечает заданное количество наиболее вероятных слов в каждой теме, присваивая им положительный или отрицательный вес на основе словаря. Следует обратить внимание, что в теме могут присутствовать слова как с положительной, так и с отрицательной тональностью. Исходя из этого, рекомендуется выгрузить списки топ-слов для всех тем вместе с тональными оценками в формат 'csv' и провести суммирование положительных и отрицательных оценок в Excel или другой сходной программе. Пользователь может решить использовать две сводные тональные оценки для каждой темы – одну, характеризующую наличие в ней позитивного сентимента, другую – негативного; либо, пользователь может усреднить эти две оценки и составить сводный индекс негативности-позитивности для каждой темы. По этим оценкам можно судить о проблематичности данной темы в глазах пользователей социальных сетей.

При использовании любых средств сентимент-анализа для русского языка следует помнить, что пока для него никому не удается добиться высокого качества. Качество определения негативно окрашенных текстов в нашим словарем около 55%, и это самый высокий результат из существующих для социально-политических текстов. Это связано и с тематикой текстов, и с особенностями русского языка.

Рекомендации по определению отношений к этническим группам

Важно понимать, что на данном этапе нет средств вычленить отношения авторов текстов к каждой из упоминаемых этнических групп в отдельности. Это связано с тем, что примерно в половине текстов упоминается более одной группе, и иногда отношение к ним противоположное. Пока не разработаны алгоритмы, способные устанавливать связь между этнонимом и смыслами, относящимися именно к нему. Однако есть ряд средств, позволяющих косвенно оценить уровень напряжения вокруг этнонима.

Во-первых, это описанный выше сентимент-анализ тем и текстов во внешнем ПО. Если среди всех текстов, в которых упоминается данная этническая группа, доля негативных текстов намного выше, чем среди текстов с другими этнонимами, это может служить признаком проблемы. То же можно сказать, если, будучи стабильно высокой, она резко изменилась в большую сторону.

Во-вторых, это функция выявления негативно окрашенных текстов с помощью алгоритма классификации внутри нашей онлайновой системы: отличие в том, что он специально обучен видеть негативный сентимент в этнорелевантных текстах.

В-третьих, это функция выявления текстов, в которых упоминается этнический конфликт. Если доля текстов с упоминанием конфликта выше среди текстов, упоминающих определенную этническую группу, это также может служить индикатором проблемы, особенно если эта доля резко выросла.

Следует понимать, что все эти методы, хотя и соответствуют уровню лучших инструментов для русского языка, имеют довольно большую погрешность. Их рекомендуется использовать совместно.

Рекомендации по визуализации результатов.

Важная задача мониторинга – отслеживание тенденций во времени и в пространстве. Для этого в нашей системе есть функции визуализации результатов тематического моделирования. Для визуализации на карте программа складывает вероятности выбранной темы в текстах, написанных пользователя из данного субъекта РФ, и получает вес интенсивности обсуждения этой темы в этом регионе. Так делается для каждого региона, на основе функции унификации географических названий; затем регионы раскрашиваются в разные цвета, в зависимости от этого веса. Это очень удобно для нахождения точек напряжения.

Вероятности тем можно суммировать и для всех текстов, написанных в данном временном промежутке (неделя, месяц, квартал, год), и расположить их на шкале времени. Это даст возможность увидеть, как менялось обсуждение выбранной темы, а пики будут индикаторами проблем, если сама тема проблемная.

Что делать после завершения всех процедур.

После завершения автоматических процедур и лейбелинга тем анализ может быть продолжен. Во-первых, разнообразная описательная статистика может многое сказать аналитику. Преобладание позитивных или негативных тем, тем об одних этнических группах над другими, распределение их по категориям пользователей (например, половозрастным), по источникам, времени и месту – все это может быть ценной информацией. Во-вторых, с некоторой аккуратностью можно применять статистический анализ; его применение ограничено ненормальным характером распределения вероятностей тем по большинству параметров (словам, текстам и как следствие группам текстов – региональным, временным и др.). Однако распределение тональности по текстам или по темам может быть нормальным или сводимым к нормальному. В-третьих, можно проводить качественный анализ текстов из интересующих тем, так как списки наиболее вероятных текстов уже представляют собой готовые выборки.

Заключение

Информационная система достаточно проста в использовании, однако все же перед использованием рекомендуется изучить руководство пользователя как к самой системе, так и к дополнительным программным средствам. В случае возникновения вопросов, комментариев можно написать письмо разработчикам на адрес: linis.hse@gmail.com