epitweetr: Пользовательская документация

Russian translation of the following document: **epitweetr : user documentation**

Описание

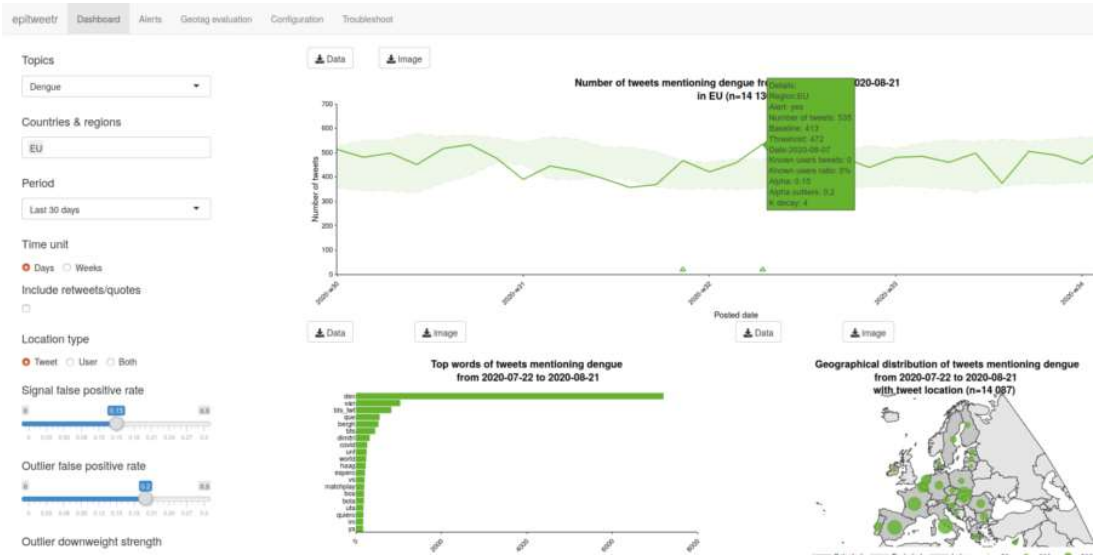
Пакет `epitweetr` позволяет вести автоматический мониторинг трендов в твитах по времени, местоположению и теме. Цель автоматического мониторинга – раннее обнаружение угроз посредством определения сигналов (напр., необычного увеличения количества твитов по определенному времени, местоположению и теме). Пакет `epitweetr` был создан с упором на инфекционные заболевания и он может быть расширен на все опасности или другие области изучения путем модификации темы и ключевых слов.

Основной принцип `epitweetr` заключается в сборе твитов и связанных метаданных из версии 1.1 стандартного поискового API-интерфейса *Twitter* (https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview/standard) в соответствии с конкретными темами и сохраняет эти твиты в сжатом виде на компьютере. Пакет `epitweetr` указывает геолокацию твитов и собирает информацию по ключевым словам внутри твита. Агрегирование твитов ведется по теме и географическому положению. Далее алгоритм обнаружения сигналов определяет количество твитов (по теме и географическому положению), которое превышает ожидаемое количество для данного дня. Затем `epitweetr` отправляет оповещения по эл. почте тем лицам, которые будут проводить дальнейшее исследование этих сигналов, руководствуясь процессом эпидемического анализа (фильтрация, проверка достоверности, анализ и предварительная оценка).

Пакет включает в себя интерактивное веб-приложение (*Shiny app*) с пятью страницами: *dashboard*, где пользователь может визуализировать и исследовать твиты (Рис. 1), страница *alerts*, где вы можете просматривать имеющиеся оповещения и связанную информацию (Рис. 2), страницу *geotag evaluation*, где вы можете оценивать алгоритм геолокации в разных полях твита, чтобы вручную выбрать порог геолокации (Рис. 3), страница *configuration*, где вы можете изменить настройки и проверить статус основных процессов (Рис. 4), и страница *troubleshoot* с автоматическими проверками и подсказками по использованию `epitweetr` и всех его функций (Рис. 5). На странице *dashboard* пользователи могут просматривать агрегированное количество твитов по их времени, расположению на карте и словам, которые наиболее часто встречаются в этих твитах. Эти визуализации могут быть отфильтрованы по теме, местоположению и временному периоду, которые вам необходимы. Доступны и другие фильтры, которые дают возможность настроить единицу времени на временной шкале, включение ретвитов/цитирования, интересующие типы геолокации, сенситивность интервала прогнозирования для обнаружения сигналов и количество дней для расчета порогового уровня сигналов. Эта информация также доступна для скачивания прямо из интерфейса в виде данных, изображений и отчетов.

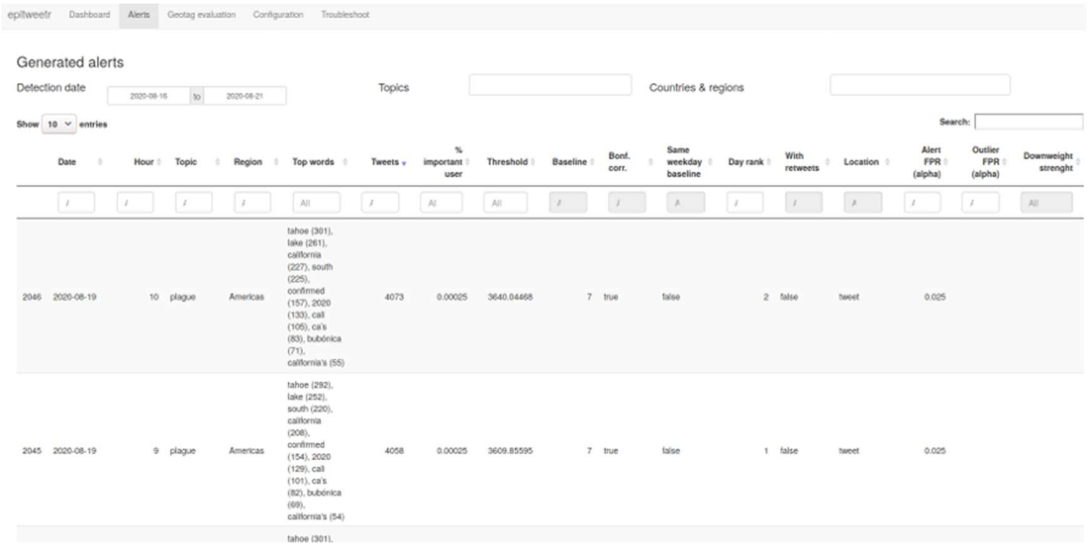
***Shiny app dashboard*:**

Рис. 1: Изображение *Shiny app dashboard*



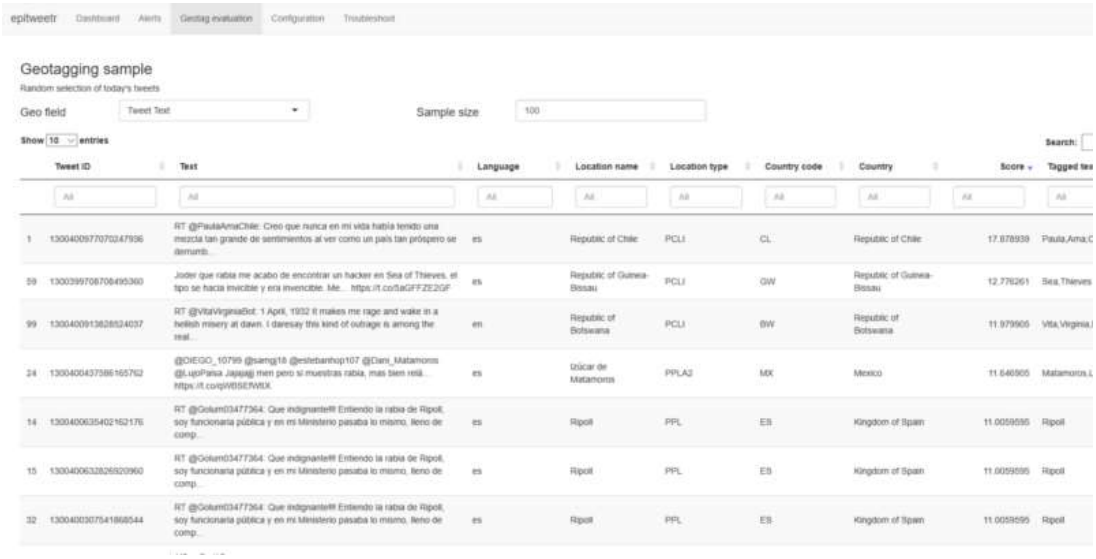
Страница *Shiny app alerts*:

Рис. 2: Страница *Shiny app alerts*



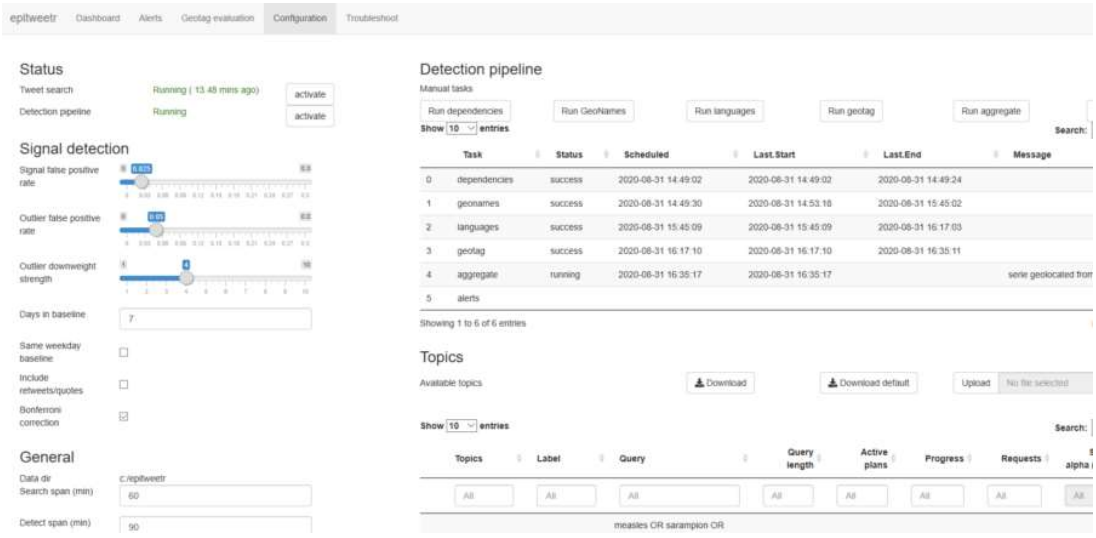
Страница *Shiny app geotag evaluation*:

Рис. 3: Страница *Shiny app geotag evaluation*



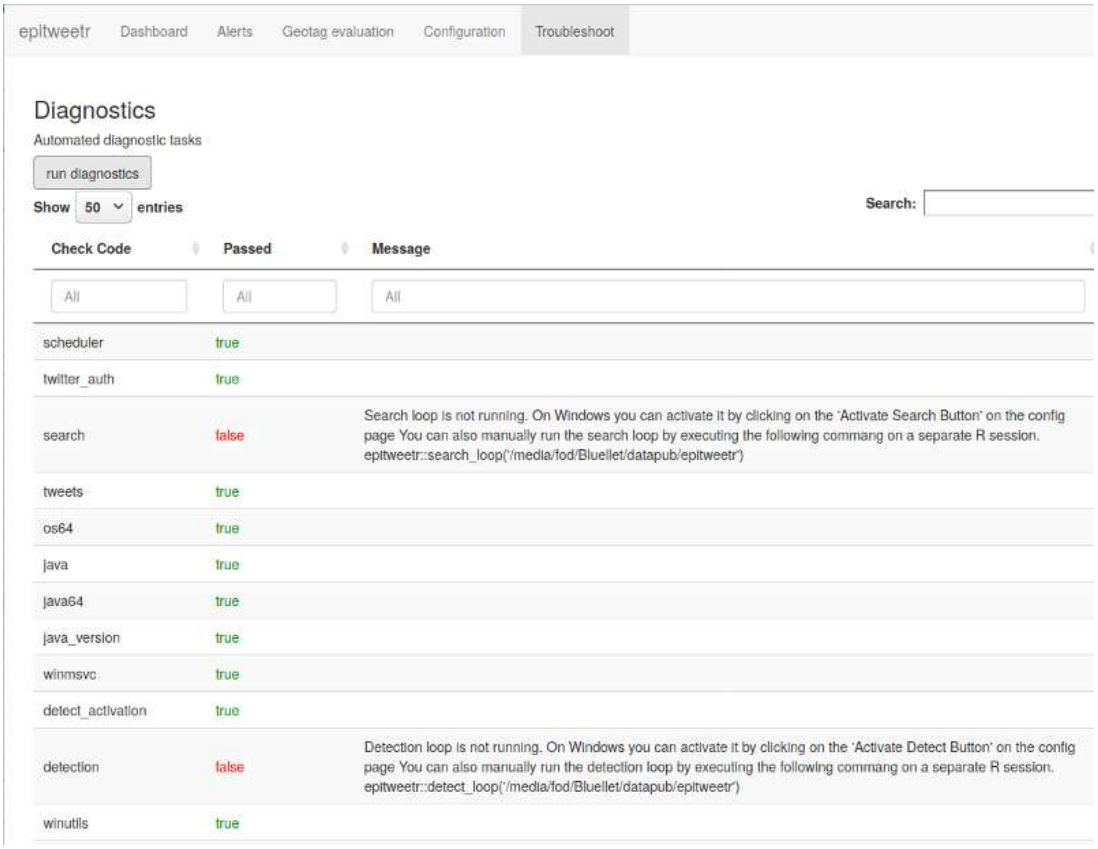
Страница *Shiny app configuration*:

Рис. 4: Страница *Shiny app configuration*



Страница *Shiny app troubleshoot*:

Рис. 5: Страница *Shiny app troubleshoot*



Общие сведения

Эпидемический анализ в ЕЦПКЗ

Статья 3 Положения о финансировании Европейского центра профилактики и контроля заболеваний (ЕЦПКЗ) и Решение № 1082/2013/ЕС о серьезных трансграничных угрозах здоровью определили обнаружение угроз общественному здоровью в качестве основного направления деятельности ЕЦПКЗ.

ЕЦПКЗ занимается эпидемическим анализом (далее – ЭА) с целью быстрого обнаружения и оценки угроз общественному здоровью, сосредотачиваясь на инфекционных заболеваниях, для обеспечения безопасности здоровья в ЕС. ЕЦПКЗ использует социальные медиа среди своих источников для раннего обнаружения сигналов об угрозах общественному здоровью. До 2020 года мониторинг социальных медиа выполнялся путем проверки и анализа сообщений заранее выбранных экспертов и организаций в основном из *Twitter* и *Facebook*.

Больше информации и онлайн-пособие доступно:

Источники ЭА

Пособие по ЭА

Мониторинг трендов в социальных медиа

Некоторые сигналы не обнаруживаются вообще или не обнаруживаются достаточно рано с помощью методов, описанных выше. Автоматический мониторинг метаданных из социальных медиа (напр., анализ трендов в социальных медиа) позволяет обнаружить сигналы, которые могут не быть обнаружены путем мониторинга заранее выбранных учетных записей в социальных медиа, и улучшает своевременность обнаружения сигналов.

Анализ трендов в социальных медиа по теме, времени и местоположению создает соответствующие сигналы для раннего обнаружения.

В 2019 году ЕЦПКЗ разработал прототип ручного инструмента на основе *R* для раннего обнаружение угроз общественному здоровью на основе данных *Twitter*. Пакет `epitweetr` – это расширение этого прототипа для обеспечения более широкой геолокации твитов и большей автоматизации.

Цели *epitweetr*

Главная цель `epitweetr` – использование версии 1.1 стандартного поискового API-интерфейса *Twitter* с целью обнаружения ранних сигналов потенциальных угроз по теме и географическим единицам.

Вторичная цель – дать возможность пользователю посредством интерактивного интерфейса *Shiny* исследовать тренды твитов по времени, географическому положению и теме, включая информацию о самых популярных словах и количестве твитов от надежных пользователей, используя диаграммы и таблицы.

Требования к оборудованию

Минимальные и рекомендуемые требования к оборудованию для компьютеров представлены в таблице ниже:

Требования к оборудованию	Минимум	Рекомендуемые
Необходимая *RAM*	8 Гб	Рекомендовано 16 Гб
Необходимая *CPU*	4 ядра	12 ядер
Объем места, необходимого для трехлетнего хранения	3 ТБ	5 ТБ

Использование *CPU* и *RAM* может быть настроено на странице *Shiny app configuration* (см. *The interactive user application (Shiny app)>The configuration page).* Необходимые *RAM, CPU* и место могут зависеть от объема и размера тем, запрашиваемых в ходе процесса сбора.

Установка

`epitweetr` создан независимым от использующихся платформ и работает на *Windows, Linux* и *Mac*. Мы рекомендуем использовать `epitweetr` на компьютере, который может работать непрерывно. Вы можете выключить компьютер, но при этом вы можете упустить некоторые твиты, если время выключения достаточно велико, что повлияет на обнаружение оповещений. Перед использованием `epitweetr`, необходимо установить следующее:

Необходимые условия для работы epitweetr

- Версия *R* 3.6.3 или выше
- *Java 1.8 eg. openjdk* версия «1.8» *https://www.java.com/download/*. Рекомендуется 64-битная, а не 32-битная версия из-за ограничений памяти. На компьютере *Mac* вам также понадобится *Java Development Kit* *https://docs.oracle.com/javase/9/install/installation-jdk-and-jre-macos.htm*]
- Если вы используете *Windows*, вам также понадобится *Microsoft Visual C++*, однако в большинстве случаев он уже будет установлен:
  - *Microsoft Visual C++ 2010 Redistributable Package (x64)* *https://www.microsoft.com/en-us/download/details.aspx?id=14632*

Необходимые условия для некоторых функциональных возможностей *epitweetr*

- *Pandoc* для экспорта файлов *PDF* и *Markdown*
  - *https://pandoc.org/installing.html*
- Установка *Tex* (*TinyTeX* или *MiKTeX*) (или другая система *TeX*) для экспорта файлов *PDF*
  - Самое простое: *https://yihui.org/tinytex/* установка из *R*, после установки требуется выход из/вход в сеанс
  - *https://miktex.org/download* требуется полная установка, после установки требуется выход из/вход в сеанс
- Оптимизация машинного обучения (только для опытных пользователей)
  - *Open Blas* (оптимизатор *BLAS*), который ускорит некоторые процессы геолокации: *https://www.openblas.net/* Инструкции по установке: *https://github.com/fommil/netlib-Java*
  - **или** *Intel MKL* (*https://software.intel.com/content/www/us/en/develop/tools/math-kernel-library/choose-download.html*)
- Планировщик

- При использовании *Windows*, необходимо установить пакет *R: taskscheduleR*
- При использовании *Linux*, необходимо планировать задачи вручную
- При использовании *Mac*, вам необходимо установить пакет *R: cronR*

Дополнительные необходимые условия для разработчиков *R*

Если вы хотите стать разработчиком `epitweetr`, необходимы следующие инструменты разработки:

- *Git* (контроль исходного кода) *https://git-scm.com/downloads*
- *Sbt* (компилятор кода *scala*) *https://www.scala-sbt.org/download.html*
- При использовании *Windows* вам также понадобится *Rtools*: *https://cran.r-project.org/bin/windows/Rtools/*

Внешние зависимости

`epitweetr` потребуется загрузить некоторые зависимости для работы. Инструмент сделает это автоматически при первом запуске процесса обнаружения оповещений. Страница конфигурации приложения *Shiny app* позволит изменить целевые *URL*-адреса этих зависимостей:

- *CRAN JAR:* Транзитивные зависимости для *Spark, Lucene* и встроенного кода *scala*. *[https://repo1.maven.org/maven2]*
- *Winutils.exe* (только для *Windows*) Это двоичный путь *Hadoop*, необходимый для локального запуска *SPARK* в *Windows [http://public-repo-1.hortonworks.com/hdp-win-alpha/winutils.exe]*.

Установка *epitweetr* из *CRAN*

После установки всех необходимых зависимостей, указанных в разделе «Необходимые условия для работы *epitweetr*», вы можете установить `epitweetr`:

```
install.packages(epitweetr)
```

Переменные среды

Кроме того, среда *R* должна знать, где установлена *Java*. Чтобы это проверить, введите в консоли *R*:

```
Sys.getenv("JAVA_HOME")
```

Если команда возвращает *NULL* или пусто, вам необходимо установить переменную среды *Java Home* для вашей ОС, см. Инструкции для конкретной ОС. В некоторых случаях `epitweetr` может работать без установки переменной среды *Java Home*.

При первом запуске приложения если инструмент не может определить надежное хранилище паролей, предоставленное операционной системой, вы увидите всплывающее окно с запросом пароля в связке ключей (*Linux* и *Mac*). Пароль необходим для хранения зашифрованных учетных данных *Twitter*. Выберите надежный пароль и запомните его. Этот пароль необходимо вводить каждый раз при запуске инструмента. Чтобы избежать этого, установите переменную среды системы под названием *ecdc_twitter_tool_kr_password*, содержащую выбранный пароль.

Запуск epitweetr Shiny app

Вы можете запустить `epitweetr` *Shiny app* из сессии *R*, печатая в консоли *R*. Замените «*data_dir*» на назначенную директорию данных, которая представляет собой локальную папку, которую вы выбираете для хранения твитов, временных рядов и файлов конфигурации в:

```
library(epitweetr)
epitweetr_app("data_dir")
```

Обратите внимание, что директория данных, введенная в *R*, должна иметь знак «/» вместо компьютерных кавычек «"»» (пример правильного пути: *'C:/user/name/Documents'*). Это особенно относится к *Windows*, если вы копируете путь из Проводника.

В качестве альтернативного варианта вы можете использовать меню запуска: В исполняемом файле *.bat* или файле оболочки введите следующие: (замените «*data_dir*» на назначенную директорию данных)

R –vanilla -e epitweetr::epitweetr_app("data_dir")

Вы можете проверить, все ли требования установлены правильно на странице диагностики *troubleshoot*. Больше информации доступно в разделе *The interactive user application (Shiny app)>Dashboard:The interactive user interface for visualisation>The troubleshoot page*

Настройка сбора твитов и цикла обнаружения оповещений

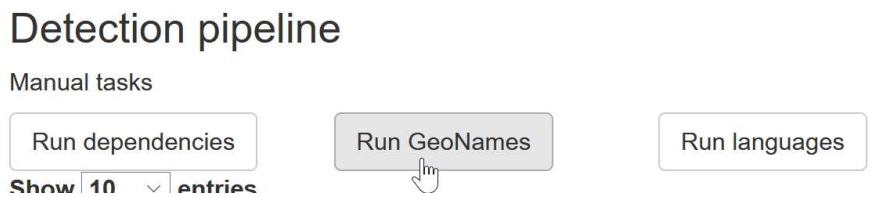
Чтобы использовать `epitweetr`, вам необходимо собирать твиты и запускать цикл обнаружения оповещений (географические имена, языки, геотеги, агрегирование и оповещения). Более подробная информация также доступна в следующих разделах пользовательской документации. Краткая информация о необходимых шагах:

- Запустите *Shiny app* (из консоли *R*)

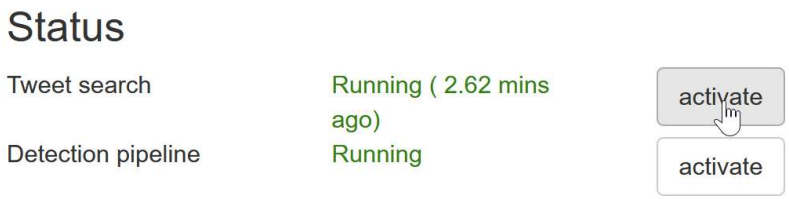
```
library(epitweetr)
epitweetr_app("data_dir")
```

- На странице конфигурации *Shiny app*, в ручных задачах «*Detection pipeline*», нажмите на «*Run dependencies*», «*Run geonames*» и «*Run languages*» (их статус изменится на «*pending*»). Это позволяет каналу обнаружения загрузить

необходимые элементы. Пока не добавлены языки и нет обновлений на *geonames.org*, эти задачи запускаются только при первой установке `epitweetr`.



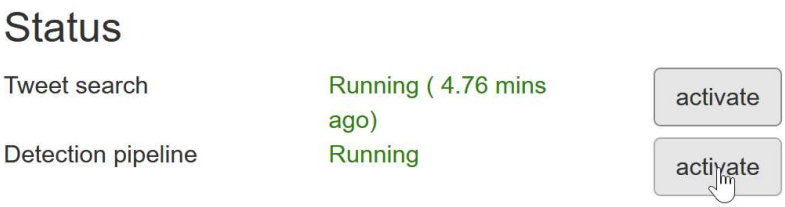
- Настройте аутентификацию в *Twitter* с помощью учетной записи *Twitter* или приложения для разработчиков *Twitter*, см. раздел *Collection of tweets>Twitter authentication* для более подробной информации
- Активировать сбор твитов
  - *Windows:* Нажмите на кнопку активации «*Tweet search*» – «*activate*»



  - Другие платформы: в новой сессии *R* выполните следующую команду

```
library(epitweetr)
search_loop("data_dir")
```

- Вы можете убедиться в том, что сбор твитов работает, если статус «*Tweet search*» *status* показывает «*Running*» на странице *Shiny app configuration* (зеленый текст на скриншоте выше) и «*true*» на странице *Shiny app troubleshoot*.
- Активировать канал обнаружения:
  - *Windows:* Нажмите на кнопку активации «*Detection pipeline*» – «*activate*»



  - Другие платформы: в новой сессии *R* выполните следующую команду

```
library(epitweetr)
detect_loop("data_dir")
```

- Вы можете убедиться в том, что канал обнаружения работает, если статус «*Detection pipeline*» *status* показывает «*Running*» на странице *Shiny app configuration* и «*true*» на странице *Shiny app troubleshoot*.
- Вы сможете визуализировать твиты после завершения шага агрегирования в таблице канала обнаружения на странице *Shiny app configuration* и если активирован «*Tweet search*».
- Вы можете начинать работу со сгенерированными сигналами. **Удачного обнаружения сигналов!**

Для большей информации перейдите к разделу *How does it work? General architecture behind* `epitweetr`, который описывает процессы, лежащие в основе сбора твитов и обнаружения сигналов. В разделе «*The interactive Shiny application (Shiny app)>The configuration page*» также описаны различные настройки на странице конфигурации.

Как это работает? Общая архитектура *epitweetr*

В следующих разделах подробно описаны общие принципы, изложенные выше. Параметры многих из этих элементов могут быть настроены на странице конфигурации *Shiny app*, что объясняется в разделе *The interactive Shiny application (Shiny app)>The configuration page*.

Сбор твитов

Использование версии 1.1 стандартного поискового API-интерфейса Twitter

`epitweetr` использует версию 1.1 стандартного поискового API-интерфейса *Twitter* Преимущество этого API-интерфейса в том, что это бесплатная услуга, предоставляемая *Twitter* и дающая пользователям `epitweetr` бесплатный доступ к твитам. Поисковой API-интерфейс не является исчерпывающим источником твитов. Он ведет поиск в выборке недавних твитов, опубликованных за последние 7 дней, и фокусируется на релевантности, а не полноте. Это значит, что некоторые твиты и пользователи могут быть пропущены в результатах поиска.

Несмотря на то, что это может быть ограничением в других сферах публичного здоровья и исследований, команда разработчиков `epitweetr` считает, что с целью обнаружения сигналов достаточно выборки твитов для определения потенциально значимых угроз в сочетании с другими типами источников.

Другие особенности версии 1.1 стандартного поискового API-интерфейса *Twitter*:

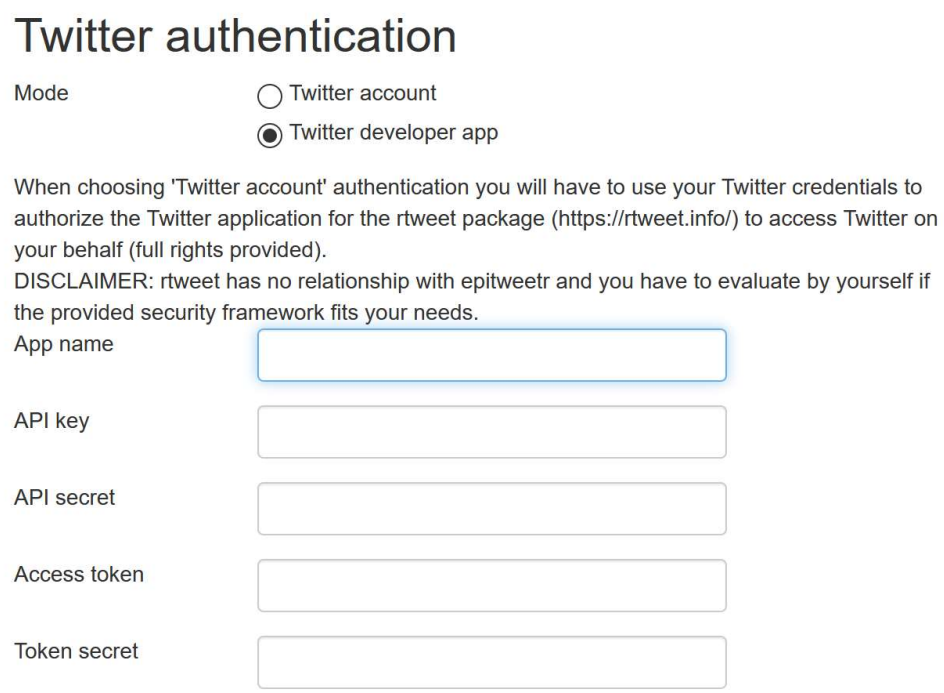
- *Twitter* индексирует только твиты за последние 5–8 дней.
- Стандартный поисковой API-интерфейс *Twitter* поддерживает максимум 180 запросов каждые 15 минут (450 запросов каждые 15 минут, если вы используете учетные данные приложения разработчика *Twitter*; см. следующий раздел)

- Каждый запрос возвращает максимум 100 твитов и/или ретвитов.

Аутентификация Twitter

Вы можете аутентифицировать коллекцию твитов, используя ***Twitter account*** (этот подход использует приложение пакета *rtweet*) или используя ***Twitter application***. Для последнего вам понадобится *Twitter developer account*, получение которого может занять некоторое время из-за процедур верификации. Мы рекомендуем использовать учетную запись *Twitter* через пакет *rtweet* для тестирования и краткосрочного использования и приложение разработчика *Twitter* – для долгосрочного использования.

- **Использование *Twitter account*:** через *rtweet* (аутентификация пользователя)
  - Вам понадобится учетная запись *Twitter* (имя пользователя и пароль)
  - Пакет *rtweet* отправит запрос в *Twitter*, чтобы получить доступ к вашей учетной записи *Twitter* от вашего имени
  - Появится всплывающее окно, в котором вам необходимо ввести свое имя пользователя и пароль в *Twitter* для того, чтобы разрешить приложению доступ к *Twitter* от вашего имени. Вы будете отправлять этот токен каждый раз для доступа к твитам.
- **Использование *Twitter developer app*:** через `epitweetr` (аутентификация приложения)
  - Если вы еще этого не сделали, вам необходимо создать учетную запись разработчика *Twitter*: *https://developer.twitter.com/en/apply-for-access*
  - Создать приложение
  - Для данного типа доступа убедитесь, что у вас есть доступ для чтения и записи.
  - Запишите ваши настройки *OAuth*
    - Добавьте их на страницу конфигурации в приложении *Shiny* (см. рис. ниже).
    - С помощью этой информации `epitweetr` может запрашивать токен в любой момент напрямую у *Twitter*. Преимущество этого метода в том, что токен не связан с какой-либо пользовательской информацией и твиты возвращаются независимо от любого пользовательского контекста.
    - С помощью этого приложения вы можете выполнять 450 запросов каждые 15 минут вместо 180 запросов каждые 15 минут, которые позволяет учетная запись *Twitter*.

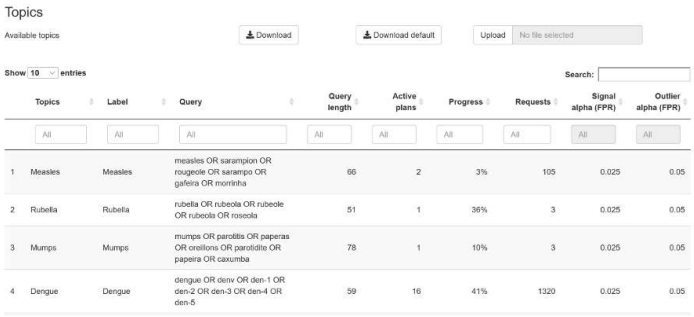


Темы и запросы на сбор твитов

После аутентификации в *Twitter* вам необходимо указать список тем в `epitweetr` для выбора, какие твиты собирать. Для каждой темы у вас есть один или более запросов, которые `epitweetr` использует для сбора соответствующих твитов (напр., несколько запросов по теме с использованием различной терминологии и/или языков).

Запрос состоит из ключевых слов и операторов, которые используются для сравнения атрибутов твитов. Ключевые слова, разделенные пробелом, обозначают оператор *AND*. Вы также можете использовать оператор *OR*. Знак минус перед ключевым словом (без пробела между знаком и ключевым словом) указывает, что ключевое слово не должно содержаться в атрибутах твита. Несмотря на то, что длина запросов может составлять до 512 символов, наилучшая практика – ограничить ваш запрос до 10 ключевых слов и операторов и ограничить сложность запроса, это значит, что иногда вам потребуется более одного запроса на одну тему.

`epitweetr` доступен со списком тем по умолчанию, которые использовались группой по эпидемическому анализу ЕЦПКБ на дату создания пакета (1 сентября 2020 г.). Вы можете просмотреть подробную информацию о списке тем на странице конфигурации приложения *Shiny app* (см. скриншот ниже).


Topics

Available topics

Download | Download default | Upload | No file selected

Show 10 entries | Search:

	Topics	Label	Query	Query length	Active plans	Progress	Requests	Signal alpha (FPR)	Outlier alpha (FPR)
	All	All	All	All	All	All	All	All	All
1	Measles	Measles	measles OR sarampion OR rougeole OR sarampo OR gafeira OR morrinha	66	2	3%	105	0.025	0.05
2	Rubella	Rubella	rubella OR rubeola OR rubeole OR rubeola OR roseola	51	1	36%	3	0.025	0.05
3	Mumps	Mumps	mumps OR parotitis OR paperas OR oreillons OR parotidite OR papeira OR caxumba	78	1	10%	3	0.025	0.05
4	Dengue	Dengue	dengue OR denv OR den-1 OR den-2 OR den-3 OR den-4 OR den-5	59	16	41%	1320	0.025	0.05

На странице конфигурации вы также можете скачать список тем, изменить его и загрузить в `epitweetr`. Новый список тем будет использоваться для сбора твитов и отображаться в приложении *Shiny app*. Список тем – это файл *Excel* (**.xlsx*), так как он хорошо справляется с региональными пользовательскими настройками (напр., разделители) и специальными символами. Вы можете создать свой собственный список тем и загрузить его, учитывая, что структура должна включать как минимум:

- Название темы с заголовком «*Topic*» в таблице *Excel*. Название должно включать только буквенно-цифровые символы, пробелы, дефисы и подчеркивания. Обратите внимание, что оно должно начинаться с буквы.

- Запрос с заголовком «*Query*» в таблице *Excel*. Это запрос, который `epitweetr` использует в своих запросах для получения твитов из стандартного поискового API-интерфейса *Twitter*. Смотрите выше информацию о синтаксисе и ограничении запросов.

Файл *topics.xlsx* дополнительно включает следующие поля:

- *ID* с заголовком «#» в таблице *Excel*, который указывает текущий целочисленный идентификатор темы.

- Ярлык с заголовком «*Label*» в таблице *Excel* обозначает то, что отображается в выпадающем меню тем на вкладках приложения *Shiny app*.

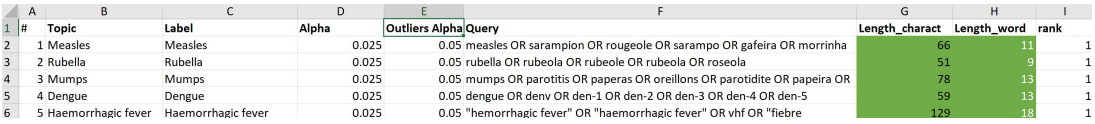
- Параметр альфа с заголовком «*Signal alpha (FPR)*» в таблице *Excel*. Сокращение *FPR* обозначает уровень ложноположительных результатов – «*false positive rate*». Повышение параметра альфа снизит порог обнаружения сигнала, что приведет к повышенной чувствительности и, возможно, получению большего количества сигналов. Установить параметр альфа можно эмпирически и в соответствии с важностью и характером темы.

- «*Length_charact*» – это автоматически генерируемое поле, которое рассчитывает длину всех символов, используемых в запросе. Это поле полезно, так как длина запроса не должна превышать 500 символов.

- «*Length_word*» указывает на количество слов, используемых в запросе, включая операторы. Рекомендуется ограничить количество ключевых слов до 10.

- Альфа-параметр с заголовком «*Outlier alpha (FPR)*» в таблице *Excel*. Сокращение *FPR* обозначает уровень ложноположительных результатов – «*false positive rate*». Этот параметр альфа устанавливает уровень ложноположительныых результатов для определения выброса при подавлении предыдущих выбросов/сигналов. Чем ниже значение, тем меньше предыдущих выбросов может быть потенциально включено. Более высокое значение потенциально может включать больше предыдущих выбросов.

- «*Rank*» – это количество запросов по теме



#	Topic	Label	Alpha	Outliers Alpha	Query	Length_charact	Length_word	rank
1	Measles	Measles	0.025	0.05	measles OR sarampion OR rougeole OR sarampo OR gafeira OR morrinha	66	11	1
2	Rubella	Rubella	0.025	0.05	rubella OR rubeola OR rubeole OR rubeola OR roseola	51	9	1
3	Mumps	Mumps	0.025	0.05	mumps OR parotitis OR paperas OR oreillons OR parotidite OR papeira OR	78	13	1
4	Dengue	Dengue	0.025	0.05	dengue OR denv OR den-1 OR den-2 OR den-3 OR den-4 OR den-5	59	13	1
5	Haemorrhagic fever	Haemorrhagic fever	0.025	0.05	"hemorrhagic fever" OR "haemorrhagic fever" OR vhf OR "fiebre	129	18	1

При загрузке своего файла измените поля «тема» и «запрос», но не меняйте заголовки столбцов.

Запланированный график сбора твитов

В качестве напоминания `epitweetr` запланирован на 180 запросов, отправляемых в *Twitter* каждые 15 минут (или 450 запросов каждые 15 минут, если вы используете учетные данные приложения разработчика *Twitter*). Каждый запрос может вернуть 100 твитов. Запросы возвращают твиты и ретвиты. Они возвращаются в легком формате передачи данных *JSON*.

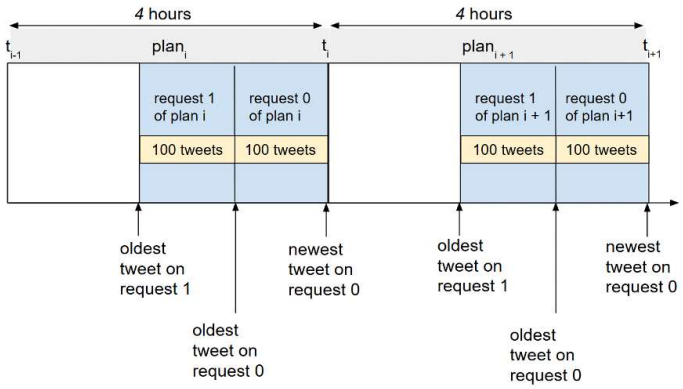
С целью сбора максимального количества твитов с учетом ограничений стандартного поискового API-интерфейса, а также для того, чтобы популярные темы не мешали достаточному сбору других тем, `epitweetr` использует «поисковые планы» для каждого запроса.

Первый «поисковой план» для запроса будет собирать твиты с текущей даты-времени на 7 дней назад (7 дней из-за ограничения стандартного поискового API-интерфейса) до момента внедрения текущего «поискового плана». Первый «поисковой план» самый большой, так как до него не было собрано никаких твитов.

Все последующие «поисковые планы» – это запланированные интервалы, которые настраиваются на странице конфигурации приложения *Shiny app* `epitweetr` (см. раздел *The interactive Shiny app > the configuration page > General*). В качестве иллюстрации будем считать, что поисковые планы запланированы с интервалами в четыре часа. Планы собирают твиты по конкретному запросу с текущей даты-времени на четыре часа назад до даты-времени внедрения текущего «поискового плана» (см. рис. ниже). `epitweetr` будет посылать столько запросов (каждый возвращает до 100 твитов) за четырехчасовой интервал, сколько необходимо для получения всех твитов, созданных в течение этого четырехчасового интервала.

Например, если «поисковой план» начинается в 16:00 10 сентября 2020 года, `epitweetr` будет запускать запросы на твиты, которые соответствуют запросам на четырехчасовой период с 4:00 до полуночи 10 сентября 2020 года. `epitweetr`

начинает со сбора самых последних твитов (с 4:00) и продолжает в обратном порядке. Если в течение четырехчасового периода с 4:00 до полуночи API-интерфейс больше не возвращает результатов, «поисковой план» для этого запроса считается выполненным.



Однако если темы очень популярны (напр., COVID-19 в 2020 году), то «поисковой план» для запроса в данном четырехчасовом окне может быть не выполнен. Если такое произойдет, `epitweetr` перейдет к «поисковым планам» для следующего четырехчасового окна и поставит все предыдущие неполные «поисковые планы» в очередь для выполнения после выполнения «поисковых планов» для нового четырехчасового окна.

Каждый «поисковой план» сохраняет следующую информацию:

Поле	Тип	Описание
expected_end	Отметка времени	Дата-время окончания текущего поискового окна
scheduled_for	Отметка времени	Запланированная Дата-время для следующего запроса. Во время создания плана это будет текущая Дата-время, а после каждого запроса это значение будет устанавливаться для будущей Даты-времени. Для установки будущей Даты-времени приложение оценит количество запросов, которые необходимо выполнить. Если оно оценит, что необходимо выполнить *N* запросов, следующее план начнется через *1/N* оставшегося времени.
start_on	Отметка времени	Дата-время завершения первого запроса плана.
end_on	Отметка времени	Дата-время завершения последнего запроса плана, если этот запрос достиг 100% выполнения плана.
max_id	*Long*	Максимальный *Twitter id*, охваченный этим планом, который будет определен после первого запроса

since_id	*Long*	Последний *id* твита, возвращенный последним запросом этого плана. Следующий запрос начнет собирать твиты до этого значения. Это значение обновляется после каждого запроса и позволяет API-интерфейсу *Twitter* возвращать твиты до *min_time(p_i)*
since_target	*Long*	Если существует предыдущий план, это значение хранит первый *id* твита, загруженного для этого плана. Текущий план не будет собирать твиты до этого *id*. Это значение позволяет API-интерфейсу *Twitter* возвращать твиты после *p_i-time_back*
requests	*Int*	Количество запросов, выполненных как часть плана
progress	*Double*	Прогресс текущего плана в процентах. Он рассчитывается как *(current$max_id - current$since_id)/(current$max_id - current$since_target).* Если API-интерфейс *Twitter* не возвращает твитов, прогресс устанавливается на 100%. Это применяется только для ответов без ошибок, содержащих пустой список твитов.

`epitweetr` будет выполнять планы в соответствии с этими правилами:

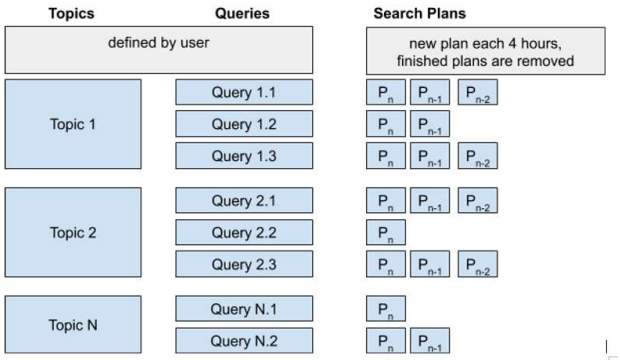
- `epitweetr` будет определять новейший незавершенный план для каждого поискового запроса с переменной *schedule_for*, расположенной в прошлом.
- `epitweetr` будет выполнять эти планы с минимальным количеством уже выполненных запросов. Это обеспечивает выполнение всеми запланированными планами одинакового количества запросов.
- В результате двух предыдущих правил запросы на темы в рамках лимита в 180 запросов для стандартного поискового API-интерфейса *Twitter* (или 450, если вы используете аутентификацию приложения разработчиков *Twitter*), будут выполняться первыми и будут показывать более значительный прогресс, чем темы, превышающие лимит.

Причина этого в том, что темы с таким большим количеством твитов, для сбора которых недостаточно четырехчасового окна поиска, скорее всего, уже являются интересующей темой. Поэтому приоритет стоит отдавать более мелким и, возможно, менее известным темам.

Примером может быть пандемия COVID-19 в 2020 году. В начале 2020 года была доступна ограниченная информация о COVID-19, которая позволяла обнаруживать сигналы со значимой информацией или обновлениями (напр., новые страны, сообщающие о случаях или подтверждающие, что они были вызваны коронавирусом). Тем не менее, в ходе пандемии эта тема стала все более популярной и широкая тема COVID-19 стала неэффективной для обнаружения сигналов и занимала много времени и запросов для `epitweetr`. В таком случае имеет смысл

установить приоритет сбора более мелких тем, таких как подтемы, связанные с COVID-19 (напр., вакцина И *COVID-19*), или убедиться, что вы не упускаете другие события, которые имеют меньше внимания в социальных сетях.

Если поисковые планы не могут быть завершены, несколько поисковых планов на запрос могут стоять в очереди:



Геолокация

Параллельно с процессом сбора твитов `epitweetr` пытается определить геолокацию всех собранных твитов, используя неконтролируемый процесс машинного обучения. Этот процесс работает в окне расписания, определенном свойством «Определить диапазон» («Detect span») на странице *configuration* в разделе «*General settings*» (напр., если установлено четырехчасовое окно, он будет запускаться каждые четыре часа и определять геолокацию всех твитов, собранных с момента последнего успешного запуска).

`epitweetr` хранит два типа геолокации для твита: местоположение твита, т.е. геолокационная информация в тексте твита (ретвита или цитируемого твита), и местоположение пользователя из доступных метаданных. Для обнаружения сигналов используется наилучшее местоположение, при этом на панели управления можно визуализировать оба типа.

Геолокация на основе местоположения твита

`epitweetr` получает и сохраняет местоположение твита на основе геолокационной информации, найденной в тексте твита. В случае ретвита или цитируемого твита геолокационная информация извлекается из оригинального текста твита, который был ретвитнут или процитирован. Если оба варианта не доступны, местоположение твита на основе текста твита не сохраняется.

`epitweetr` определяет, имеет ли текст твита ссылку на определенное местоположение путем разбивки текста твита на наборы слов и оценки тех, которые вероятнее всего могут быть местоположением при помощи модели машинного обучения. Алгоритм также добавляет последующие слова (одно за другим) к набору, и если оценка возрастает вследствие использования большего количества слов, алгоритм пытается найти локальный максимум оценки в большем тексте. Далее он

сравнивает эти слова со справочной базой данных – *geonames.org*. Это географическая база данных, открытая и доступная через различные веб-сервисы в соответствии с лицензией *Creative Commons attribution*. База данных *GeoNames.org* содержит более 25 000 000 географических названий. `epitweetr` использует по умолчанию только существующие сейчас населенные пункты и те, население которых известно (то есть чуть более 500 000 названий). Вы можете переопределить это значение на странице конфигурации приложения *Shiny app configuration*, сняв галочку с «Упрощенные геоназвания» («Simplified geonames»). База данных также содержит атрибуты долготы и широты и варианты написания (перекрестные ссылки) местностей, которые полезны для поиска целей, а также нелатинские варианты написания многих из этих названий.

Сравнения могут выполняться на любом уровне административной иерархии. Сравнение обеспечивает *Apache Lucene,* это высокопроизводительная полнофункциональная текстовая поисковая библиотека с открытым кодом.

Некоторые тексты могут иметь несколько местоположений в процессе сравнения. Тем не менее, будет выбрано только местоположение с наивысшей оценкой.

Более высокая оценка связана с большей вероятностью верного совпадения. Оценка:

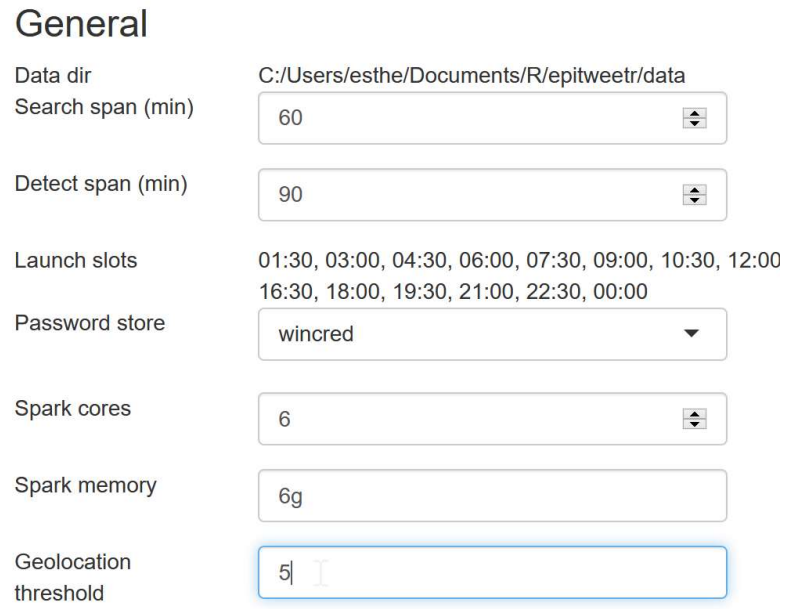
- Выше, если необычные части названия совпадают
- Выше, если несколько административных уровней совпадают
- Выше, если население местности больше
- Выше для стран и городов, чем для административных уровней
- Выше для аббревиатур из заглавных букв, напр., *NY*
- Ниже для слов, которые, скорее всего, относятся к другим (негеографическим) типам слов. Например, город «Фэир Плей» в Колорадо. Это достигается за счет использования языковых моделей *fasttext.cc*.

Вы можете выбрать, на каких языках (*languages*) вы хотите проверять другие слова, выбрав активный язык на странице *configuration* в *Shiny app* и нажав на иконку «+»:

Кроме того, вы можете отменить выбор языков, выбрав язык на странице конфигурации приложения *Shiny app* и нажав иконку «–»: *ADD IMAGE HERE*

Минимальная оценка («*geolocation threshold*») может быть установлена глобально в общих настройках на странице конфигурации для снижения количества ложноположительных результатов (см. рис.). Все геолокации с оценкой ниже, чем порог геолокации, будут отклонены алгоритмом как местоположение твита. Если существует более одного совпадения свыше минимальной оценки, то будет выбрано совпадение с самой высокой оценкой.

Порог выбирается эмпирически и может быть оценен на основе прочтения твитов человеком и местоположений твитов на странице оценки геотегов.



Геолокация на основе местоположения пользователя

Различные типы местоположений пользователей доступны из метаданных, предоставляемых стандартным поисковым API-интерфейсом *Twitter*. epitweetr выбирает лучшие местоположения пользователя для агрегации файлов в следующем порядке:

- точное или приблизительное местоположение пользователя на момент твита (предоставляется API-интерфейсом)
- если местоположение пользователя недоступно и твит – это ретвит или цитируемый твит, то используется точное или приблизительное местоположение пользователя на момент ретвита/цитируемого твита API-интерфейсом)
- если это недоступно, используется указанное пользователем местоположение
- если это недоступно, используется значение «дом» («*home*») в публичном профиле.

Наряду с точным местоположением показаны долгота и широта. Если метоположение приблизительное, `epitweetr` рассчитывает долготу и широту из *GeoNames.org*.

Если информация о местоположении пользователя, предоставляемая API-интерфейсом, недоступна, то `epitweetr` рассчитает долготу и широту на основе заявленного пользователем местоположения или названия места, указанного в «публичном профиле» пользователя, используя *GeoNames.org*.

Сохраненная геолокационная информация твитов

Геолокация совпадения сохраняется в виде кода страны (используя стандарт ISO 3166), а также долготы и широты, связанных с точной геолокацией в агрегированных данных.

Самые частотные слова в твитах

Так как количество твитов и слов может быть очень большим, твиты анализируются по наиболее частым словам в блоках по 10 000 твитов, чтобы получить топ-500 слов из каждого блока для одного и того же языка, дня и темы.

Чтобы обеспечить адекватную производительность `epitweetr`, топ-500 слов сначала определяются на глобальном уровне, затем эти слова используются для выборки по странам, чтобы получить топ-слова по стране, дню и теме. Самые частотные слова могут не встречаться в небольших населенных пунктах с малым количеством твитов, определенных геолокацией.

Учтите, что, в отличие от других визуализационных данных, самые часто встречающиеся слова в твитах всегда основаны на геолокации, связанной с «местоположением твита», а не с «местоположением пользователя», независимо от фильтра, выбранного на панели управления.

Агрегирование данных

Процесс агрегирования данных создает три файла *Rds* (родной формат *R*): *geolocated*, *country_counts* и *topwords*.

В геолокационном файле *Rds* количество твитов или ретвитов сохраняется по теме, дате, долготе и широте геолокации текста твита и долготе и широте геолокации пользователя. Каждая из этих позиций также включает страну, ассоциируемую с геолокацией текста твита, и страну, ассоциируемую с геолокацией пользователя (см. частичный скриншот ниже). Имейте в виду, что твиты без геолокационной информации также включены.

	topic	created_date	user_geo_country_code	tweet_geo_country_code	user_geo_code	tweet_geo_code	user_geo_name	tweet_geo_name
1	COVID-19	2020-08-20	IT	AR	IT	AR	Italian Republic	Argentine Republic
2	COVID-19	2020-08-20	CA	US	6113365	4521037	Prince George	Pennsville
3	COVID-19	2020-08-20	US	US	4726206	4726206	San Antonio	San Antonio
4	gonorrhea	2020-08-20	CO	CO	CO	3687459	Republic of Colombia	El Caracol

Файл *country_counts Rds* используется для создания линии тренда в *Shiny app*. Это меньший файл *Rds* без информации о долготе и широте, который включает

количество твитов по часам в течение дня, по стране (в соответствии с местоположением твита или местоположением пользователя), теме (см. скриншот) и наличием или отсутствием его ретвита. Поля *known_retweets* и *known_original* показывают количество твитов или ретвитов из списка «*important users*». В этот файл также включены твиты без геолокации. Включение твитов без геолокационной информации позволяет вам видеть все твиты при выборе «*world*» в качестве региона, независимо от успешности геолокации.

	topic	created_date	created_hour	tweet_geo_country_code	user_geo_country_code	retweets	tweets	known_retweets
33	COVID-19	2020-08-16	19	AU	US	71	13	0
34	COVID-19	2020-08-16	19	GH	PK	5	3	0
35	rabies	2020-08-16	21	PK	ES	20	0	0
36	gonorrhea	2020-08-16	01	VE	NA	1	0	0
37	COVID-19	2020-08-16	04	NA	PE	88	22	0

Агрегирование по топовым словам сохраняется в файле *topwords.Rds* и показывает количество твитов или ретвитов (или обоих) по теме, топовому слову, дате, стране расположения твита и наличия или отсутствия ретвита (см. скриншот).

	tokens	topic	created_date	tweet_geo_country_code	frequency	original	retweets	created_weeknum
85486	crisis	Zika	2020-08-17	BA	1	1	0	202034
85487	crisllenime	plague	2020-08-17	SK	1	1	0	202034
85488	crismenshova	malaria	2020-08-17	ID	4	3	1	202034
85489	crismenshova	malaria	2020-08-17	RO	4	3	1	202034
85490	crisoliveira89	seasonal%20influenza	2020-08-17	RO	3	1	2	202034
85491	crispación	rabies	2020-08-17	AZ	1	1	0	202034
85492	crisper	seasonal%20influenza	2020-08-17	BS	1	1	0	202034
85493	crispr	Ebola	2020-08-17	CN	21	3	18	202034
85494	crispr	Ebola	2020-08-17	SL	6	2	4	202034

Обнаружение сигналов

Главная цель `epitweetr` – обнаружение сигналов в наблюдаемых потоках данных, т.е. встречаемостей в агрегированных временных рядах, которые превышают ожидаемые. Для обнаружения сигналов `epitweetr` использует расширенную версию алгоритма EARS *(Early Aberration Reporting System) (Fricker, Hegler, and Dunfee 2008)*, который далее обозначается как `eears` (расширенный *EARS*). Этот алгоритм является частью пакета *R* `surveillance` (*Salmon, Schumacher, and Höhle 2016*).

По умолчанию он использует движущееся окно последних семи дней для расчета порога. Если встречаемость за текущий день превышает этот порог, генерируется сигнал.

Подробности алгоритма, лежащего в основе обнаружения сигнала

Алгоритм `eears` применяется к встречаемости сигналов за последние семь 24-часовых блоков до текущего 24-часового блока обнаружения сигнала. Вычисляется скользящее среднее и рабочее стандартное отклонение:

$$\overline{y}_0 = \frac{1}{7}\sum_{t=-7}^{-1} y_t \quad \text{и} \quad s_0^2 = \frac{1}{7-1}\sum_{t=-7}^{-1}(y_t - \overline{y}_0)^2,$$

где $y_t, t = \cdots, -2, -1{,}0$ означает временные ряды наблюдаемых счетных данных с временным индексом 0, который означает текущий блок. Более того, временной индекс $-7, \ldots, -1$ означает семь блоков до текущего блока.

Когда верна нулевая гипотеза об отсутствии всплесков, предполагается, что y_t одинаково и независимо от $N(\mu, \sigma^2)$ распределяется с неизвестным средним значением μ и неизвестным отклонением σ^2. Поэтому верхний предел простого одностороннего интервала предсказания плагина $(1 - \alpha) \times 100\%$ для y_0, основанного на $y_{-7}, \ldots, y_{-1}$, представлен как

$$U_0 = \overline{y}_0 + z_{1-a} \times s_0,$$

где z_{1-a} – это (1 – α)-квантиль стандартного нормального распределения. Оповещение возникает, если $y_0 > U_0$. При использовании α = 0,025, то это соответствует исследованию, если y_0 превышает прогноз для среднего значения плюс 1,96 стандартного отклонения. Тем не менее, как указывают *Allévius* и *Höhle* (2017), правильным подходом было бы сравнение наблюдений с верхним пределом двухстороннего 95% интервала прогнозирования для y_0, так как здесь учитывается и колебание выборки нового наблюдения, *и* неопределенность, возникающая из прогнозирования параметров среднего значения и отклонения. Поэтому статистически подходящей формой является вычисление верхнего предела так:

$$U_0 = \overline{y_0} + t_{1-a}(7-1) \times s_0 \times \sqrt{1 + \frac{1}{7}}.$$

где $t_{1-a}(k-1)$ значит 1 – α квантиль t-распределения с k – 1 степеней свободы.

Подавление предыдущих сигналов

Если предыдущие сигналы включаются без изменений в исторические значения при расчете скользящего среднего значения и стандартного отклонения для обнаружения сигнала, то прогнозируемое среднее значение и стандартное отклонение могут стать слишком большими. Это может значить, что важные текущие сигналы не будут обнаружены. Чтобы решить эту проблему, `epitweetr` подавляет предыдущие сигналы, чтобы прогноз среднего значения и стандартного отклонения был скорректирован с учетом таких выбросов, используя подход, схожий с тем, который использовал *Farrington et al*. (1996). Исторические значения, которые не определяются как предыдущие сигналы, получают весовое значение «1». Подобным образом исторические значения, которые определяются как сигналы, получают весовое значение ниже единицы, и выполняется новый поиск соответствий с использованием этих весовых значений (расширенные так, что они снова суммируются на 7 наблюдений). Подробности процедуры подавления доступны в *Annex I* данной пользовательской документации.

Регулирование времени обнаружения сигнала

Обнаружение сигналов выполняется по «дням», которые представляют собой 24-часовое движущееся окно, которое движется в соответствии с диапазоном обнаружения (см. также раздел *The interactive user application (Shiny app) > The configuration page > General*). Базовый уровень рассчитывается на основе этих «дней» от –1 до –8 (если текущий день – это ноль).

Сигналы генерируются в соответствии с диапазоном обнаружения (см. раздел *The interactive user application (Shiny app) > The configuration page > General),* с

- общими оповещениями по эл. почте после завершения диапазона обнаружения (напр., если диапазон обнаружения был четыре часа, оповещения по эл. почте будут отправляться каждые четыре часа);
- оповещения по эл. почте в режиме реального времени. В этих оповещениях по эл. почте будут пропущены ранее сгенерированные сигналы.

Различные типы оповещений по эл. почте для каждого пользователя могут быть установлены на странице конфигурации (см. раздел *The interactive user application (Shiny app) > The configuration page > General).*

Параметр альфа: уровень ложноположительного обнаружения сигналов

Главный атрибут обнаружения сигналов – это способность алгоритма обнаруживать настоящие угрозы или события без перегрузки исследователя слишком большим количеством ложноположительных сигналов. Таким образом, параметр альфа определяет порог интервала обнаружения. Если параметр альфа высокий, то генерируется больше потенциальных сигналов, если параметр альфа низкий, то генерируется меньше потенциальных сигналов (но потенциальные угрозы или события могут быть упущены). Настройка параметра альфа часто производится эмпирически и зависит также от ресурсов исследователей сигналов и важности пропуска потенциальной угрозы или события.

Существует глобальный параметр альфа, который можно установить/изменить в `epitweetr` на странице *configuration* под «*Signal false positive rate*» (см раздел *The interactive user application (Shiny app) > The configuration page > General).* Помимо этого, параметр альфа по умолчанию может быть переопределен в списке тем. Здесь при желании можно соединить каждую тему с определенным параметром альфа в зависимости от предполагаемой важности темы для общественного здоровья или потенциального связанного события или угрозы.

Поправка Бонферрони

Для учета множественного сравнения при обнаружении сигнала по конкретной стране параметр альфа делится на количество стран по умолчанию. Для обнаружения сигнала по конкретному континенту параметр альфа делится на количество континентов. Это поправка Бонферрони для множественного сравнения.

Чтобы изменить это, вы можете снять галочку с опции «Поправка Бонферрони» в части «Обнаружение сигнала» на странице конфигурации в приложении *Shiny app*.

Использование одинаковых дней недели в качестве базовой линии

Возможно существование «эффекта дня недели», когда в определенный день недели (напр., понедельник) может быть опубликовано больше твитов, чем в другие дни. Во избежание этого вы можете выбрать расчет базовой линии не для

последовательных дней, а для последних *N* дней, которые соответствуют одинаковому 24-часовому окну *N* дней назад. В таком случае, если *N* = 7, базовая линия рассчитывается с использованием «дней» из –7, –14, –21, –28, –35, –42, –49 и –56 (если текущий «день» равен нулю).

Эта опция находится на странице конфигурации приложения *Shiny app* «*Default same weekday baseline*».

Отправка оповещений по эл. почте

`epitweetr` автоматически отправляет эл. письма, содержащие список обнаруженных сигналов, в соответствии с диапазоном обнаружения и списком подписчиков. С учетом времени, необходимого для сбора, геолокации и агрегирования твитов, в оповещениях по эл. почте будут пропускаться самые последние твиты, которые еще не прошли через эти процессы. Задержка между твитами и оповещениями должна быть меньше *(2 * (collect_span) + detect_span),* что должно составлять 3,5 часа при использовании стандартных значений.

Оповещения по эл. почте будут включать следующую информацию о сигналах по каждой теме:

- Дата и время обнаружения сигнала
- Географическое(-ие) положение(-я), где был обнаружен сигнал
- Самые частые слова (топ-слова) в твитах
- Количество твитов и порог
- Процент твитов от важных пользователей
- Информация о настройках: использовалась ли поправка Бонферрони, использовалась ли базовая линия по одному и тому же дню недели, были ли включены ретвиты и пр.

Эта информация также доступна на странице оповещений приложения *Shiny app*.

Подписчики могут получать оповещения в режиме реального времени (т.е. как только завершится цикл обнаружения) или запланированные оповещения по расписанию (напр., один или два раза в день). Список подписчиков может быть изменен на странице конфигурации, загрузив таблицу *Excel*. В этом файле имеются следующие переменные:

- «*User*»: имя подписчика (напр. *Jane Doe*).
- «*Email*»: эл. почта подписчика (напр. jane.doe@email.com).
- «*Topics*»: список тем, по которым подписчик будет получать запланированные оповещения. Используемые названия должны совпадать со столбцом «Тема» в списке тем.

- «*Excluded*»: темы, по которым подписчик не будет получать запланированные оповещения.
- «*Real time Topics*»: список тем, по которым подписчик будет получать оповещения в режиме реального времени.
- «*Regions*»: список регионов, по которым подписчик будет получать запланированные оповещения.
- «*Real time Regions:*» список регионов, по которым подписчик будет получать оповещения в режиме реального времени.
- «*Alert Slots*»: это слоты цикла обнаружения, после которых подписчик будет получать запланированные оповещения. Доступные слоты берутся из «*Launch slots*» в разделе «*General*» на странице *configuration*. Если значение не включено, подписчик будет получать уведомления в режиме реального времени для всех тем и регионов, даже если в таблице *Excel* указаны темы или регионы в режиме реального времени.

При включении более одной темы и/или региона в списке подписчиков необходимо их разделить точкой с запятой (;) без пробелов (напр., Эбола;инфекционные заболевания;денге). Названия должны совпадать со столбцом «*Topics*» в списке тем и столбцом «*Name*» в списке стран/регионов на странице конфигурации.

	B	C	D	E	F	G	H	I
	User	**Email**	**Topics**	**Excluded Topics**	**Real time Topics**	**Regions**	**Real time Regions**	**Alert Slots**
	Jane Doe	jane.doe@email.com			infectious diseases;zoonoses		Southern Europe;EU+EEA	8;20

Структура папки

`epitweetr` сохраняет твиты, агрегированные твиты и настройки в папке «*data folder*», которую вы должны указать при запуске приложения.

Внутри папки ***data*** находится 3 файла *JSON*:

- *properties.json*, созданный из информации из Общих свойств приложения *Shiny app*;
- *topics.json*, управляемый циклом обнаружения: он следит за планами сбора твитов и прогрессом;
- *tasks.json*, управляемый циклом обнаружения: он хранит информацию и статус различных задач, выполняемых этим процессом.

Имеются также следующие подпапки:

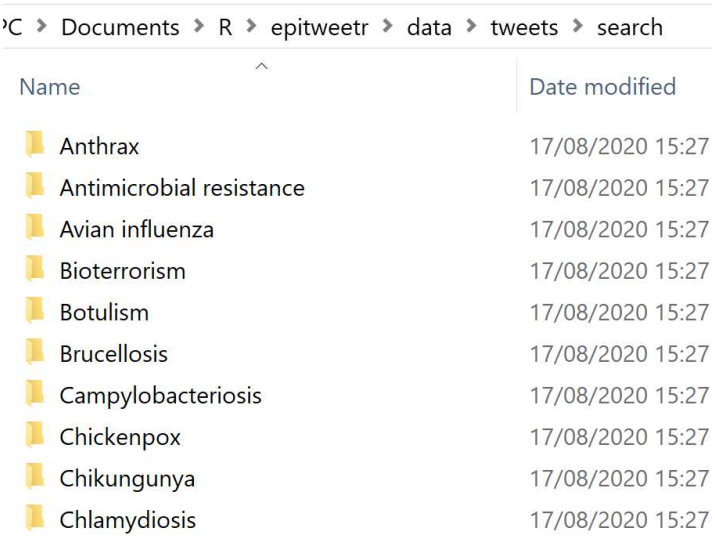
- «*geo*», где хранятся данные *GeoNames* в виде текстовых и индексных файлов
- «*hadoop*», где хранятся зависимости *Spark* для операционных систем *Windows*

- «*jars*», где хранятся коллекции зависимостей *Java*, необходимых в процессах определения геолокации и агрегирования
- «*languages*», где хранятся индексы файлов *fasttext* и модели, которые используются для определения геолокации в тексте твита
- «*stats*», где хранятся файлы *json*, докладывающие о статистике, используемой для оптимизации процесса агрегации, путем соединения файлов твитов и дат публикации твитов
- «*alerts*», которые хранят файлы *json* оповещений, полученных в цикле обнаружения.
- «*tweets*» и «*series*», которые детально описаны ниже.

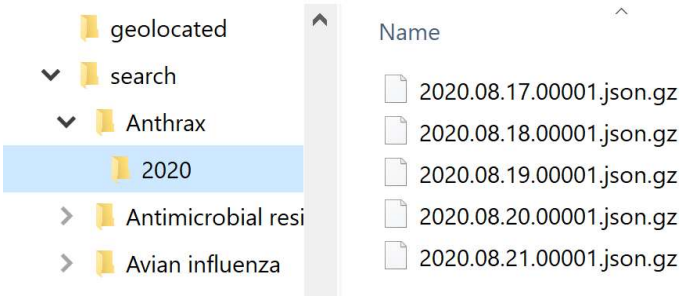
Data folder > tweets

В папке *data* подпапка «*tweets*» имеет две последующие подпапки: ***search*** и ***geolocated***

Папка search содержит подпапки для каждой темы из списка тем:



Внутри каждой темы имеется год (напр., 2020) и сжатый файл *json*, содержащий твиты по дням для каждого года. Даты относятся к датам сбора твита (не публикации). На один день может иметься более одного файла, если размер файла превышает 100 МБ.



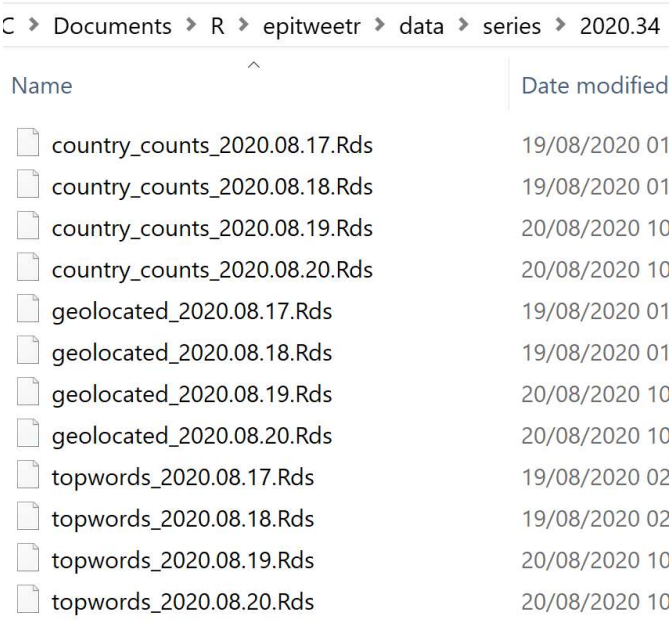
Папка geolocated содержит сжатые файлы *json* с геолокационной информацией, сгенерированной алгоритмом геолокации.

Data folder > series

В папке *series* `epitweetr` сохраняет агрегированные данные твитов с установленной геолокацией, а также топ-слова.

Существует папка для каждой недели *ISO* даты сбора, в которой содержатся файлы *Rds* (файл в родном формате *R*) для каждого дня и серии:

- *geolocated_YYY.MM.DD.Rds* содержит дневное количество твитов на максимальном уровне местоположения
- *topwords_YYY.MM.DD.Rds* содержит дневное количество топ-слов в твитах на уровне страны
- *country_counts_YYY.MM.DD.Rds* содержит часовое количество твитов на уровне страны



Это агрегированная информация, описанная в разделе «*How does it work? General architecture of epitweetr* > *Aggregation*».

Интерактивное пользовательское приложение (*Shiny app*)

Вы можете запустить интерактивное пользовательское приложение `epitweetr` (*Shiny app*) из сесии *R*, введя в консоли *R* (замените «*data_dir*» на необходимую директорию данных):

```
epitweetr_app("data_dir")
```

В качестве альтернативного варианта вы можете использовать меню запуска: В исполняемом bat или *sh* файле введите следующее: (замените «*data_dir*» на ожидаемую директорию данных)

R –vanilla -e epitweetr::epitweetr_app('data_dir')

Интерактивное пользовательское приложение `epitweetr` имеет пять страниц:

- *dashboard*, где пользователь может визуализировать и анализировать твиты
- Страница *configuration*, где вы можете изменять настройки и проверять состояние основных процессов
- Страница *alerts*, где вы можете просматривать текущие оповещения и связанную информацию
- Страница *geotag evaluation*, где вы можете оценить алгоритм геолокации в разных полях твита, чтобы вручную выбрать порог геолокации
- Страница *troubleshoot* с автоматическими проверками и подсказками по использованию `epitweetr` со всей его функциональностью

Dashboard: интерактивный пользовательский интерфейс для визуализации

Dashboard – это интерфейс, в котором вы можете интерактивно анализировать визуализацию твитов. Он включает линейный график (линию трендов) с оповещениями, картой и топовыми словами твитов по заданной теме. Учитывайте, что при первом выборе периода вам нужно **подождать 1–2 минуты, чтобы увидеть результаты**. Также при любом новом выборе (добавление региона, изменение темы и пр.), `epitweetr` начнет считывать подходящие данные; поэтому, если сделано несколько выборов, вам придется подождать 1–2 минуты, пока последний выбор не отобразится на панели управления. Когда `epitweetr` считывает новые данные, выводимые данные имеют меньшую интенсивность, указывая на то, что новые данные считываются и отображаются на графике.

Чтобы интерактивно анализировать данные, вы можете выбрать из нескольких фильтров: темы, страны и регионы, период времени, единицы времени, достоверность сигнала и дни на базовой линии.

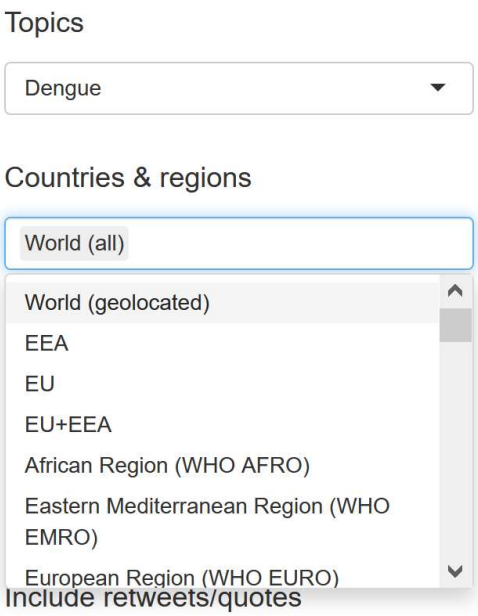
Учитывайте, что любые параметры/настройки, которые вы выбрали на панели управления, не влияют на обнаружение оповещений. Все настройки обнаружения оповещений выбираются на странице конфигурации приложения *Shiny app*.

Filters

Topics

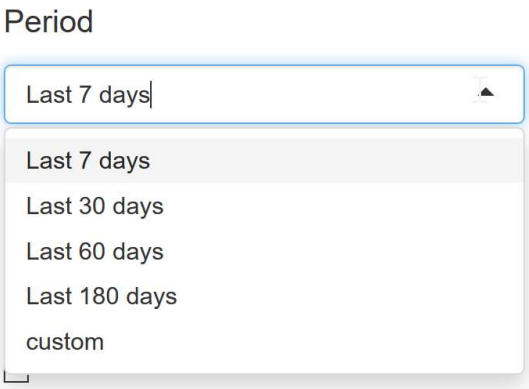
Вы можете выбрать одну из тем из выпадающего списка, который состоит из того, что указано в темах на странице конфигурации. Вы также можете начать вводить текст в текстовое поле и выбрать отфильтрованные темы из выпадающего списка.

Countries & regions



При выборе ***World (all)*** будут отображены все твиты независимо от их геолокации. Вы можете выбрать отдельную страну, регионы и субрегионы, а также несколько пунктов одновременно. Вы также можете начать вводить текст в текстовое поле и выбрать географический пункт выпадающего списка.

Period



Вы можете выбрать из последних 7 (по умолчанию), 30, 60 или 180 дней. Вы также можете выбрать «*custom*», при этом появится опция календаря для выбора периода изучения. Эти периоды будут периодом времени для включения в визуализации. При выборе собственного периода убедитесь, что первая дата выбрана минимум на один день ранее второй даты.

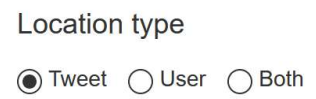
Единица времени

Вы можете показать временную шкалу для количества твитов, используя в качестве единиц времени недели или дни. По умолчанию единица времени – дни.

Include Retweets/quotes

По умолчанию ретвиты не включаются в какие-либо визуализации. Если отмечено «*include retweets/quotes*», визуализации будут показывать результаты твитов и ретвитов/цитат. В противном случае визуализации будут показывать только твиты (без ретвитов/цитат).

Location type



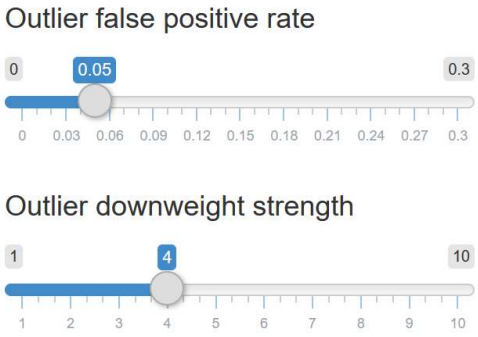
Геолокация твитов определяется по регионам, субрегионам и странам. «*Location type*» («Тип локализации») указывает, что должно использоваться для определения геолокации:

- *Tweet*: включает географическую информацию, которая содержится в тексте твита, или, если она недоступна, географическая информация, содержащаяся в тексте ретвита/цитаты, если таковая имеется.
- *User*: географическая информация, полученная из местоположения пользователя. В порядке приоритетности это местоположение пользователя на момент написания твита, местоположение API-интерфейса пользователя или пункт «дом» («*home*») в публичном профиле, если ничего из этого не доступно.
- *Both*: географическая информация для твита будет в порядке приоритетности местоположением внутри текста твита, при ее отсутствии – местоположением пользователя.

Signal detection false positive rate

С помощью слайдера вы можете анализировать различия в сигналах, генерируемых при изменении параметра альфа для уровня ложноположительных результатов. Учитывайте, что это не изменит уровень ложноположительных сигналов для эл. писем с оповещениями. Это инструмент для изучения пользователем данного параметра. По умолчанию он составляет 0,025. Более высокий уровень ложноположительных результатов повысит чувствительность и, возможно, количество обнаруженных сигналов, и наоборот.

Outlier false positive rate* и *outlier downweight strength



Outlier false positive rate относится к определению выброса при подавлении предыдущих выбросов/сигналов. Чем ниже значение, тем меньше предыдущих выбросов может быть потенциально включено. Более высокое значение потенциально может включать больше предыдущих выбросов.

Outlier downweight strength определяет, насколько будет подавлен выброс. Чем выше значение, тем больше будет подавление. Для более подробной информации см. *Annex I*.

Поправка Бонферрони

Поправка *Bonferroni correction* выбирается по умолчанию. Она учитывает обнаружение ложноположительного сигнала при множественном сравнении. Для обнаружения сигнала по конкретной стране параметр альфа делится на количество стран. Для обнаружения сигнала по конкретному континенту параметр альфа делится на количество континентов.

Если вы не хотите использовать эту поправку, снимите с ней галочку.

Days in baseline

Количество дней в базовой линии по умолчанию – 7. Пользователь может изучить влияние выбора разных дней на базовую линию. Это относится только к визуализации, любые изменения для эл. писем с оповещениями должны быть сделаны на странице конфигурации.

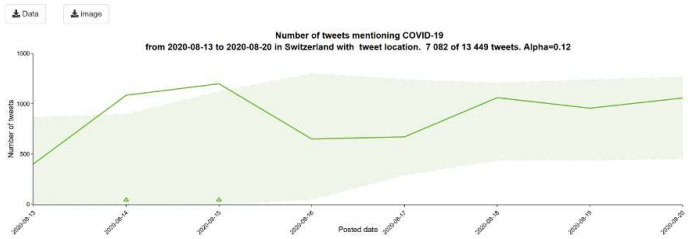
Same weekday baseline

Возможно существование «эффекта дня недели», когда в определенный день недели (напр., понедельник) может быть опубликовано больше твитов, чем в другие дни. Вы можете выбрать рассчитать базовую линию не для последовательных дней, а для последних *N* дней, которые соответствуют одинаковому 24-часовому окну *N* дней назад. В таком случае, если *N* = 7, базовая линия рассчитывается с использованием «дней» из –7, –14, –21, –28, –35, –42, –49 и –56 (если текущий «день» равен нулю).

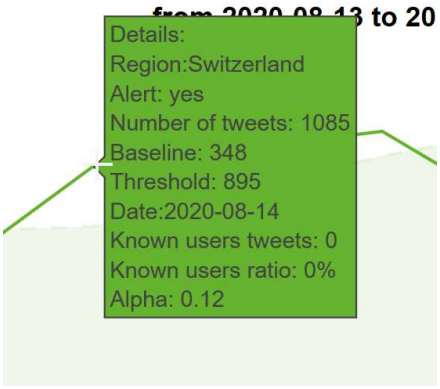
Временная линия

График временной шкалы – это временной ряд, в котором вы видите количество твитов по определенной теме, географической единице и периоду изучения. Сигналы обозначаются на графике треугольниками с указанием параметра альфа и дней базовой линии, как установлено в фильтрах. Зона ниже порога закрашена

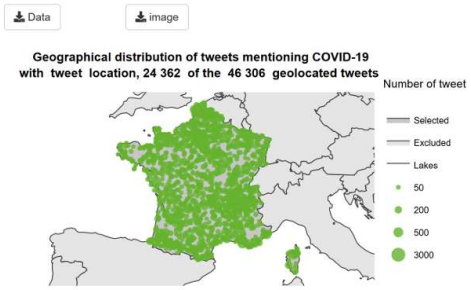
зеленым цветом. Имейте в виду, что сигналы связаны с выбором параметров альфа и дней в базовой линии в фильтрах на панели управления, а не с настройками эл. писем с оповещениями. Так вы можете анализировать влияние изменения этих параметров и адаптации настроек для эл. писем с оповещениями при необходимости.



При наведении мышкой на график вы увидите дополнительную информацию о стране, дате, количестве твитов и количестве твитов из списка известных пользователей, соотношении известных и неизвестных пользователей, количестве твитов, связанных с сигналом, и пороге параметра альфа.

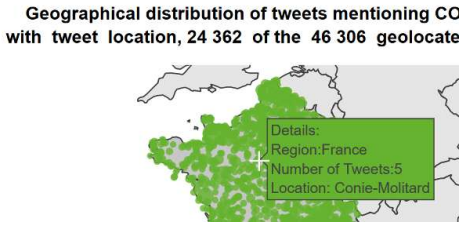


Карта



Здесь показана пропорциональная символическая карта твитов по странам и темам за анализируемый период. Чем больше кружок, тем больше количество твитов.

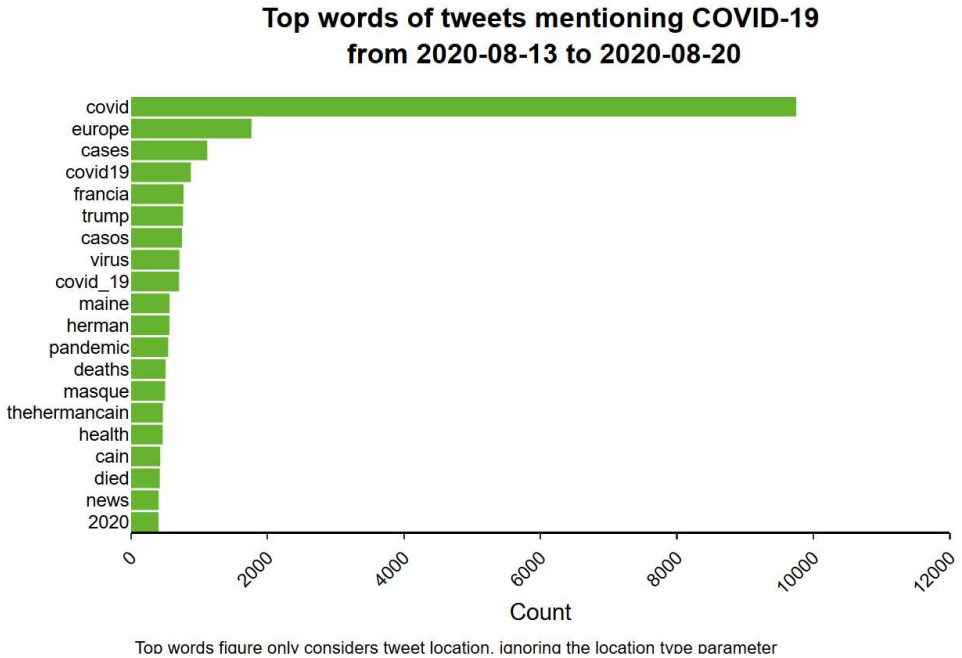
Географическая информация на карте зависит от выбора фильтров: страна/регион/субрегион и тип местоположения (твит, пользователь или оба).



При наведении мышкой на кружки на карте вы можете увидеть информацию о количестве твитов и названиях географических единиц.

При выборе одной страны символы показывают географическое распределение твитов на субнациональном уровне. При выборе двух или более стран или другой географической единицы (напр., региона или континента) символы показывают географическое распределение твитов на национальном уровне. Учтите, что если твит имеет геотег на уровне страны (напр., Франция), он не будет отображаться при выборе только этой страны, так как здесь отсутствует субнациональный геотег.

Самые частотные слова, встречающиеся в твитах



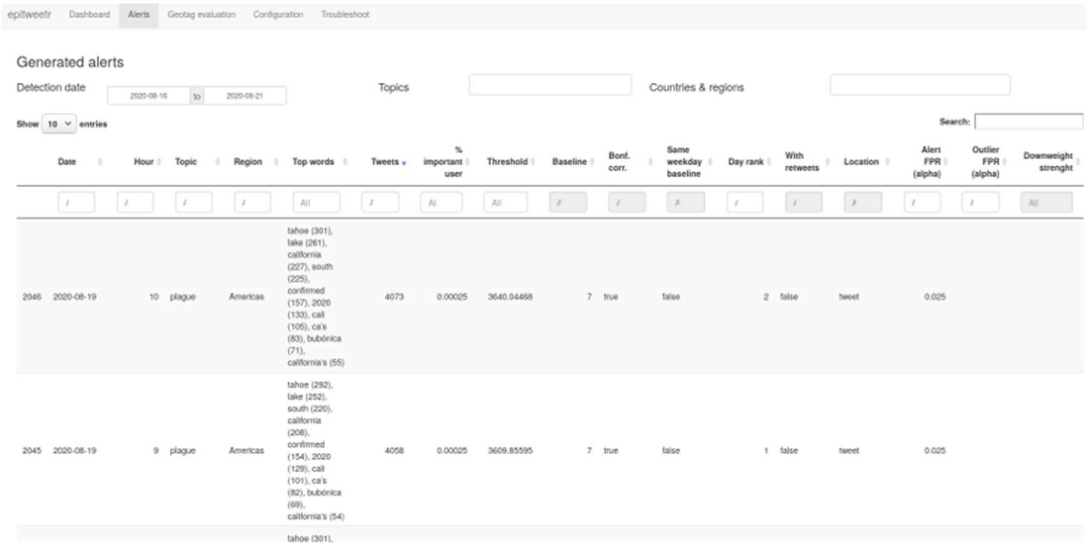
Твиты анализируются глобально для выявления первых 500 слов из каждого фрагмента для одного языка, дня и темы. Далее эти 500 слов сортируются по странам.

Далее на графике отображаются топовые слова твитов по теме за анализируемый период для выбранных географических единиц и в соответствии с фильтром твитов/ретвитов.

Учтите, что, в отличие от других визуализаций, здесь самые частотные слова в твитах всегда связаны с «местоположением твита» и на них не влияет выбор местоположения (пользователь или местоположение твита) в фильтре.

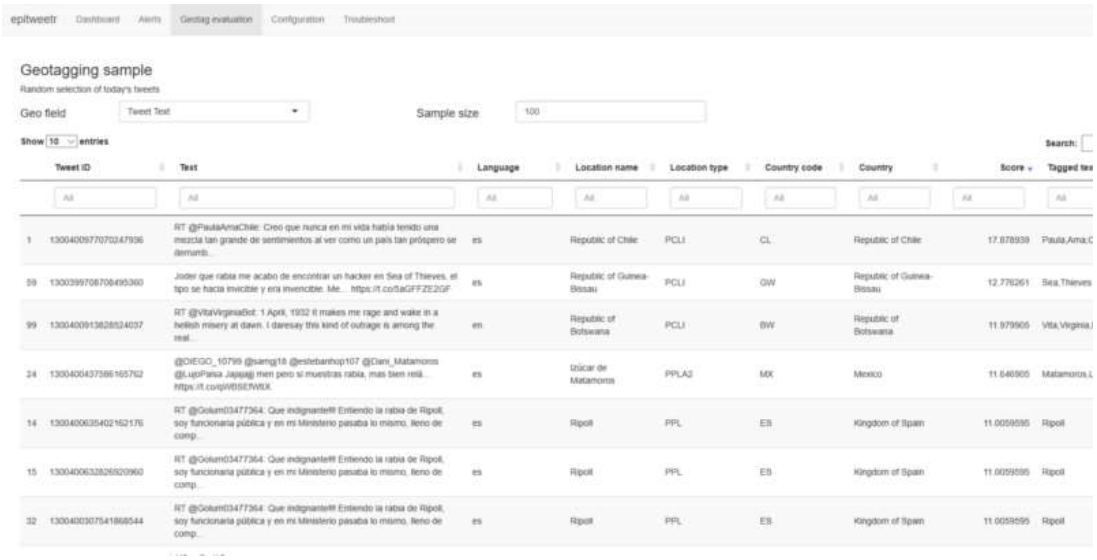
Страница *alerts*

Страница *alerts* суммирует сигналы, обнаруженные в течение конкретного анализируемого периода. Здесь включены дата, час, тема и географическая единица сигнала, топовые слова, количество твитов, количество твитов от важных пользователей и порог. Здесь также имеются многие настройки, указанные на странице конфигурации, которые были использованы для обнаружения сигналов. Эти результаты также отправляются в оповещениях по эл. почте.



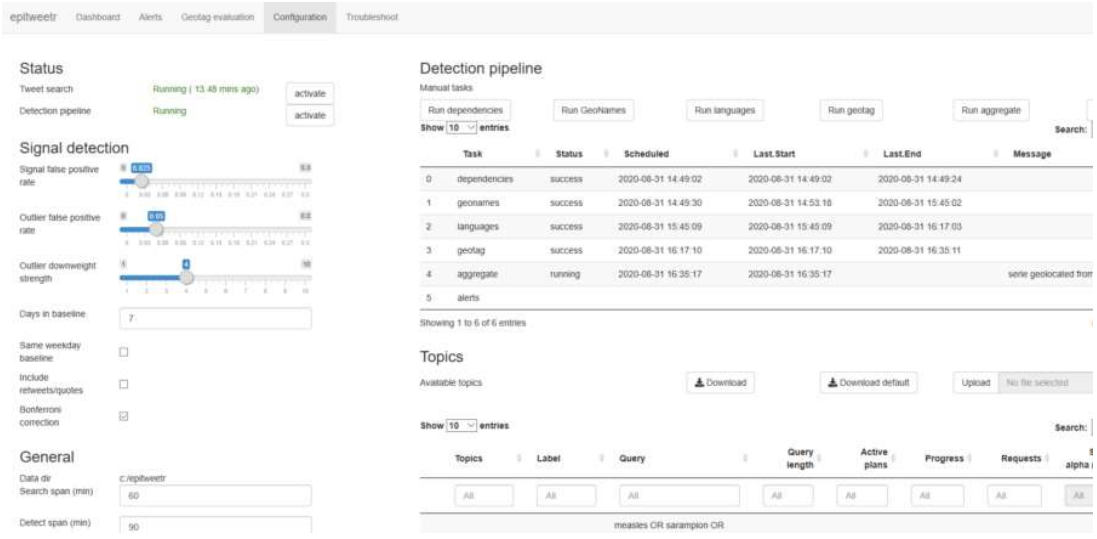
Страница *geotag evaluation*

Эта страница позволяет пользователю установить порог геолокации на странице конфигурации. Пользователь может выбрать поле твитов для тестирования и количество твитов для отбора. Эта страница используется только для визуализации, здесь недоступны никакие изменения в геолокации, обнаруженных `epitweetr`.



Страница *configuration*

На странице *configuration* вы можете изменить настройки инструмента, проверить статус различных процессов/каналов инструмента, а также добавлять, удалять и менять темы и связанные с ними запросы, языки для геолокации, список «важных пользователей» и подписчики на эл. письма с оповещениями. При изменении чего-либо в разделах «*Signal detection*» или «*General*» не забывайте нажать кнопку «*Update Properties*» в конце раздела «*General*». В следующих разделах будет более подробно описана страница *configuration*.



Status

Раздел *status* позволяет быстро оценить последний момент времени и/или статус процессов для сбора твитов (*Tweet Search*), геолокации, агрегации и обнаружения сигналов (*Detection pipeline*).

В разделе *status* вы можете увидеть, работают ли процессы канала поиска и канала обнаружения. При работе с *Windows* вы можете нажать «*activate*» и зарегистрировать эти процессы как запланированные задачи и запускать их вручную из *Windows task scheduler*.

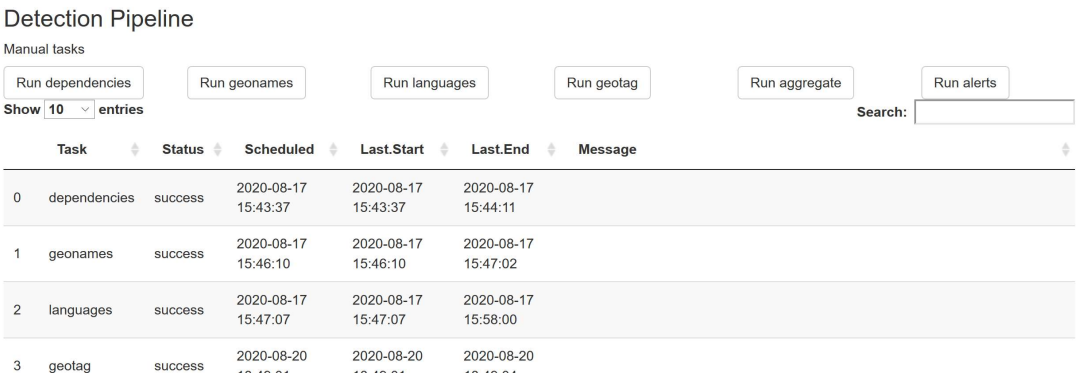
Канал обнаружения

В ручных задачах в разделе «*Detection pipeline*» вам необходимо вручную запускать задачи зависимостей, геоназваний и языков, нажав кнопки «*Run dependencies*», «*Run geonames*» и «*Run languages*» при первом использовании `epitweetr` и далее только при скачивании новых версий. Геоназвания и языки связаны с геолокационными и языковыми моделями, используемыми *epitweetr*. Если вы желаете их обновить (нерегулярно, приблизительно раз в год), вы можете нажать «*Run*».

Кнопки «*Run geotag*», «*Run aggregate*» и «*Run alerts*» могут использоваться для принудительного запуска этих задач в случае ошибки или проблемы. Вы можете проверить их статус в таблице «*Detection Pipeline*».

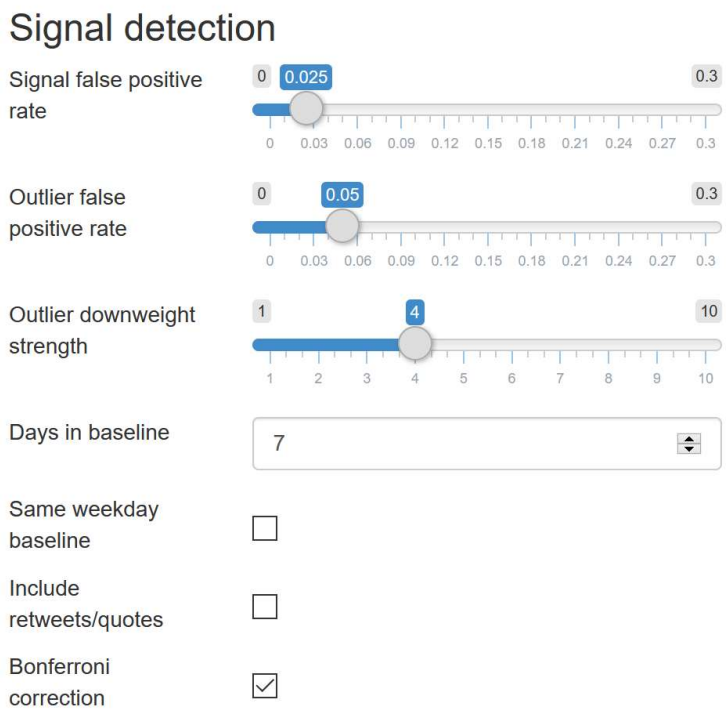
Канал обнаружения предоставляет больше информации о статусе процессов `epitweetr`. Он необходим для устранения неполадок и отслеживания прогресса. Он содержит пять задач, которые выполняются в фоновом режиме. Геоназвания и языки – это задачи, которые будут скачивать и обновлять их локальные копии. Данный процесс инициируется только в случае добавления языка или обновления *GeoNames*. Даты начала и завершения обычно намного более ранние, чем даты геотегов, агрегирования и оповещений.

Даты геотегов, агрегирования и оповещений должны быть более поздними, если каналы поиска и обнаружения активны и работают. Они запланированы в соответствии с диапазоном обнаружения. *Status* может включать *running, scheduled, pending, failed* или *aborted* (если сбой случался более трех раз).



Signal detection

В разделе *signal detection* на странице *configuration* вы можете настроить параметр альфа для уровня ложноположительного сигнала, который повышает (если больше) интервал обнаружения (обнаруживается больше сигналов) или снижает (если меньше) интервал обнаружения (обнаруживается меньше сигналов).



Outlier false positive rate относится к определению выброса при подавлении предыдущих выбросов/сигналов. Чем ниже значение, тем меньше предыдущих выбросов может быть потенциально включено. Более высокое значение потенциально может включать больше предыдущих выбросов.

Outlier downweight strength определяет, насколько будет подавлен выброс. Чем выше значение, тем больше будет подавление. Для более подробной информации см. *Annex I*.

`epitweetr` рассчитывает порог, чтобы определить, превышает ли текущее количество твитов для данного 24-часового окна ожидаемое (см. раздел «*How does it work? General architecture behind epitweetr > Signal detection*»). Этот порог основан на значениях по умолчанию за предыдущие 7 дней. В поле «дни по умолчанию в базовой линии» вы можете изменить количество дней.

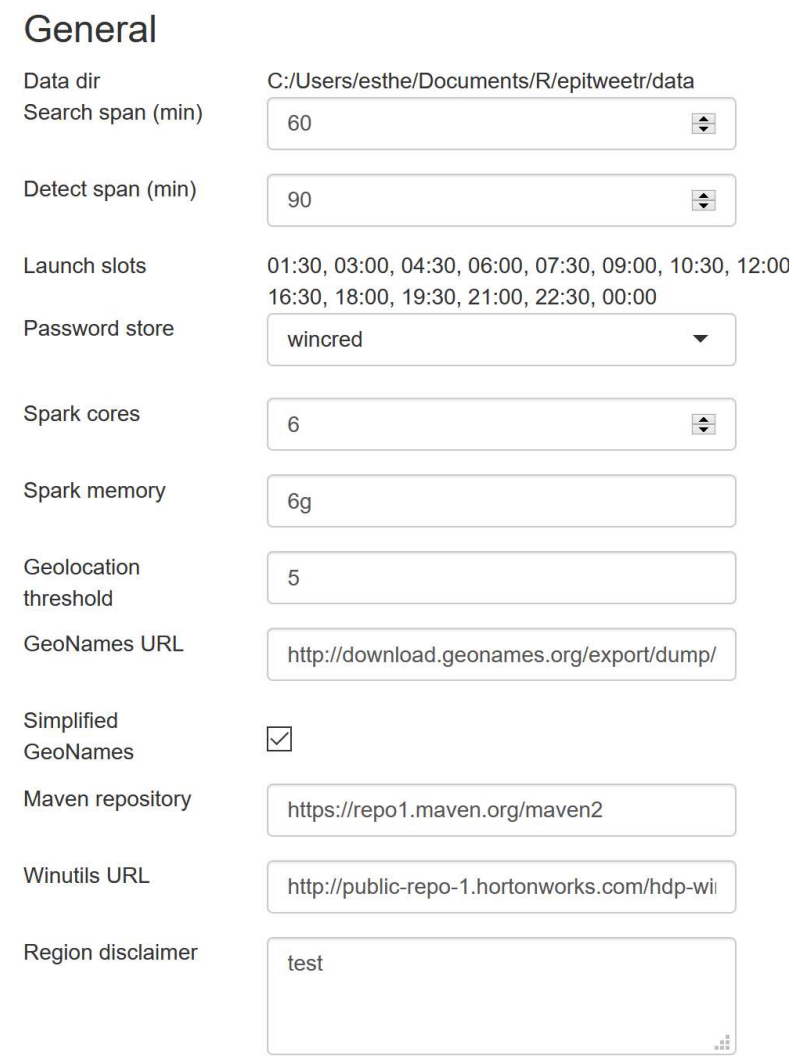
Вы также можете изменить использование 7 предыдущих дней по умолчанию для расчета базовой линии на 7 одинаковых дней недели, чтобы избежать «эффекта дня недели» (может быть так, что, например, в понедельник всегда больше твитов на эту тему, что может влиять на обнаружение сигнала).

Вы можете также установить, как выполняется обнаружение сигнала: только по тексту твита или включая ретвиты/цитаты (поставьте галочку «*Default with retweets/quotes*»).

Последняя опция «*Default with Bonferroni correction*» учитывает множественное сравнение, которое может привести к ложноположительным результатам. Если эта опция отмечена галочкой, то параметр альфа обнаружения сигнала делится на количество географических положений, в которых ведется обнаружение сигнала. Например, на уровне стран параметр альфа делится на общее количество стран. Например, на континентов параметр альфа делится на общее количество континентов.

При изменении чего-либо в разделах «*Signal detection*» не забывайте нажать кнопку «*Update Properties*» в конце раздела «*General*».

General



- В ***Data directory*** вы можете просматривать директорию, которую `epitweetr` использует для сохранения собранных данных твитов и связанных данных. Это также каталог, который панель управления использует для получения наборов данных для показа визуализаций. Вам необходимо выбрать эту папку при запуске `epitweetr` или установить переменную среды *'EPI_HOME'*.
- ***Search span*** относится к продолжительности выполнения плана поиска. По умолчанию он составляет 60 минут. Это значение контролирует размер поискового окна твитов. При уменьшении этого значения вы получите твиты раньше, но вы можете пропустить запросы на темы с малым количеством твитов. При повышении этого значения вам потребуется больше времени, чтобы получить твиты, но вы также получите больше запросов на популярные твиты, что может повысить вероятность их исчерпываемости. Вы уведите, что у

вас нет возможности собирать твиты на странице конфигурации приложения *Shiny* при наличии более, чем одного активного плана для некоторых тем.

- ***Detect span*** относится к частоте выполнения процессов канала обнаружения (геотеги, агрегирование и обнаружение оповещений). По умолчанию он составляет 90 минут. Оповещения по эл. почте отправляются в конце цикла обнаружения. Это значение считается нижней границей, цикл обнаружения может занять больше времени в зависимости от объема твитов и системных спецификаций.

- **Launch slots** для процессов канала обнаружения будут расположены в соответствии с «*Detect span*», где первый начинается в полночь. Эти значения можно использовать в файле подписчиков на странице конфигурации.

- Чтобы избежать сохранения учетных данных *Twitter* в простых файлах, `epitweetr` использует системно-зависимую функцию хранения паролей, которая хранится в ***Password store***. В зависимости от вашей системы вы можете выбрать механизм, который подходит для среды работы `epitweetr`. Подробности о каждом вводе см. *https://CRAN.R-project.org/package=keyring*

  - *wincred:* (только *Windows*) использует диспетчер учетных данных *Windows*.

  - *macos:* (только *MAC*) использует службы связки ключей *Mac OS*

  - *file:* Использует зашифрованные файлы, защищенные паролем

  - *secret service:* (только *Linux*) использует секретную службу *Linux*

  - *environment:* Использует переменные среды (необходима дополнительная настройка, см. *https://CRAN.R-project.org/package=keyring*)

- ***Spark cores* и *spark memory*:** Выделение памяти для `epitweetr` *CPU* (ядра *Spark*) и *RAM* (память *Spark*) также определяется в разделе «*general*». По умолчанию выделено 6 ядер и 6 Гб *RAM*. Это будет зависеть от объема *CPU* и *RAM* вашего компьютера и должно быть равно или меньше упомянутого.

- ***Geolocation threshold:*** Во время процесса геолокации наборы слов обрабатываются в поисках потенциальных совпадений с существующими местоположениями, которым присваивается оценки. Чем выше оценка, тем выше вероятность, что геолокация верна. Порог устанавливается в `epitweetr`, ниже которого любые совпадения считаются недостаточными для геолокации. Диапазон шкалы – от 1 до 10, по умолчанию установлено 5.

- ***Geonames URL:*** *URL*-адрес, используемый для скачивания базы данных *GeoNames* (для генерирования местоположений) находится в разделе «Общее». При изменении этого *URL*-адреса вы можете внести изменения здесь.

- ***Simplified geonames:*** Так как *GeoNames* – это очень большой файл, его упрощенная версия используется по умолчанию и включает только существующие географические пункты, население которых известно. Вы можете снять галочку с этой опции, если хотите использовать всю базу данных *GeoNames*.
- ***Maven repository:*** Это *URL*-адрес репозитория *maven*, который будет использоваться для загрузки зависимостей *JAR* для цикла обнаружения, в основном *Spark* и *Lucene*.
- ***Winutils URL:*** Это *URL*-адрес, который будет использоваться для загрузки *winutils.exe*. Это двоичный файл *Windows* для локального запуска *Spark* в *Windows*. Если вы не желаете использовать эту версию, вы можете создать ее сами, загрузив *Hadoop 2.8.4* или выше и скомпилировав ее в *Windows*.
- ***Region disclaimer:*** Если вы желаете добавить заметку к используемой карте. Заметка добавляется к экспортированному изображению карты панели управления, а также к экспортированному *PDF*-файлу панели управления.

Twitter authentication

У вас есть два варианта аутентификации для сбора твитов: используя учетную запись *Twitter* (используя пакет *rtweet*) и используя приложение для разработчиков *Twitter*. Вы можете выбрать вариант аутентификации в разделе ***Twitter authentication***. См. раздел «*How does it work? General architecture behind epitweetr > Collection of tweets > Twitter authentication"* для подробной информации об аутентификации *Twitter*.

Email authentication (SMTP)

В данном разделе необходимо указать детали эл. почты для *SMTP* аутентификации для эл. почты, с которой будут отправляться оповещения.

Если ***Unsafe certificates*** отмечено галочкой, то `epitweetr` будет использовать ваш *SMTP*-сервер, даже если сервер отправляет недействительный сертификат.

При изменении чего-либо в разделе «*general*» не забывайте нажать кнопку «*Update Properties*».

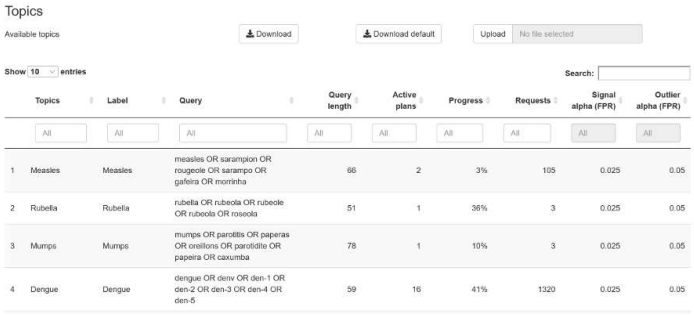
Topics

Темы определяют, какие твиты собирает `epitweetr`. Это делается в таблице *Excel*, которая содержит темы и связанные запросы, которые `epitweetr` использует для отправки запроса в *Twitter API*-интерфейс.

Запрос состоит из ключевых слов и операторов, которые используются для сравнения атрибутов твитов. См. раздел «*How does it work? General architecture behind epitweetr > Collection of tweets > Topics of tweets to collect and queries*» для подробной информации о запросах.

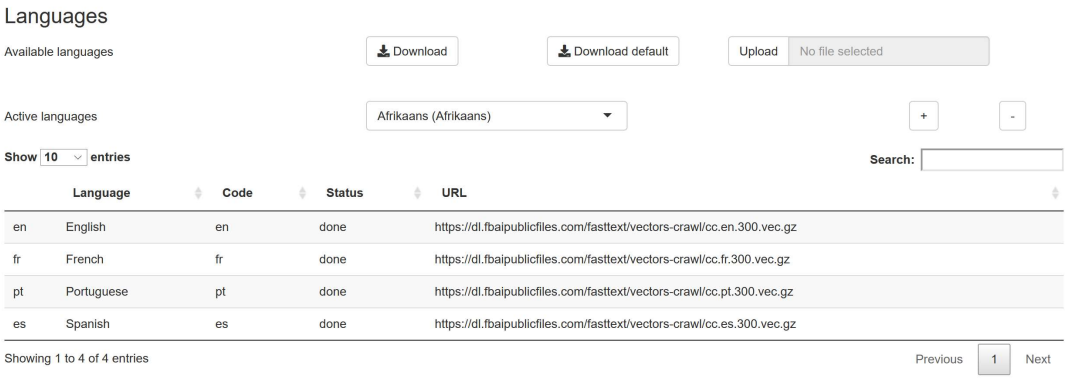
`epitweetr` доступен со списком тем по умолчанию, которые использовались группой по эпидемическому анализу ЕЦПКБ на дату создания пакета (1 сентября 2020 г.). Вы можете скачать список тем и загрузить свой собственный в раздел «*Available Topics*» на странице *configuration*. См. раздел «*How does it work? General architecture behind epitweetr > Collection of tweets > Topics of tweets to collect and queries*» для подробной информации о структуре списка тем.

В разделе тем на странице конфигурации вы можете видеть тему, связанный запрос, длину запроса и количество активных планов поиска, связанных с запросом. Если активен более, чем один план, это значит, что `epitweetr` не смог собрать все возможные твиты за предыдущую сессию. Кроме этого, вы можете видеть прогресс и количество запросов из предыдущего поискового плана.



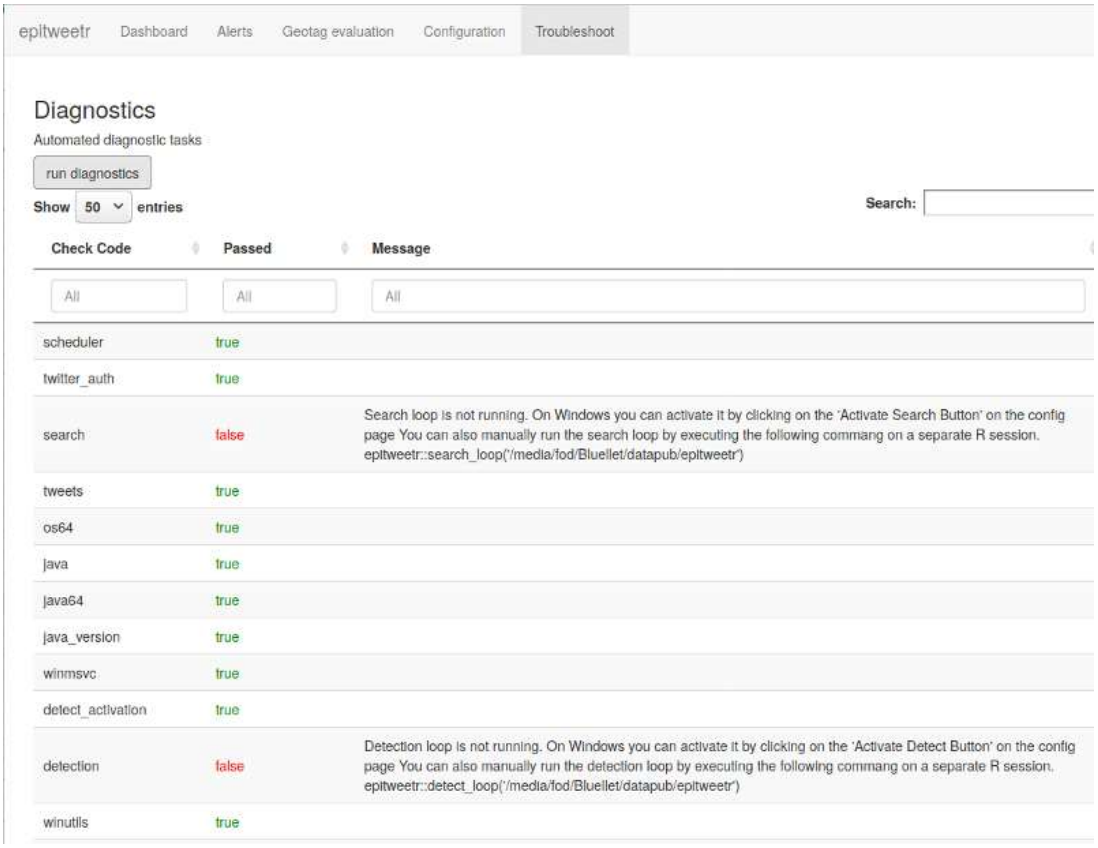
Languages

В разделе языков вы можете установить, какие языковые модели будут использоваться для идентификации текста в процессе геолокации. По умолчанию используются французский, английский, португальский и испанский языки. Вы можете скачать и загрузить языковые модели в разделе «*Available Languages*» и добавить и удалить языки, используемые `epitweetr` в разделе «*Active Languages*». Учитывайте вычислительные мощности для добавления слишком большого количества языков в зависимости от мощности вашего компьютера.



Страница *troubleshoot*

Страница *troubleshoot* имеет список автоматических проверок и подсказок по использованию `epitweetr` со всей его функциональностью. Нажмите на «*Run diagnostics*», чтобы увидеть, какие проверки выполнены («*true*»), а какие нет («*false*»), и подсказки в случае непрохождения проверки. Более подробная информация доступна в приложении Annex II данной пользовательской документации.



Скачивание выводимых данных из интерактивного пользовательского интерфейса (*Shiny app*)

Каждая визуализация на панели управления *Shiny app* может быть скачана в виде изображения, используя кнопку «*image*». *png* – это портативный сетевой графический файл и универсальный формат для изображений, разрешение которых может быть не очень высоким (напр., профессиональная графика для типографии).

Учтите, что формат *png* не поддерживается в браузере *Internet Explorer* (вместо него вы можете скачать файл *svg*).

Вы также можете скачать данные каждой визуализации, нажав на кнопку данных. Вы получите файл *csv* с исходными данными, которые вы можете использовать для дальнейшего анализа или создания собственных графиков.

В качестве альтернативы вы можете использовать *PDF*-файл или кнопку *Md* внизу фильтров, чтобы скачать *PDF* или *HTML*-файл панели управления. Учтите, что для этого вам потребуется установленный *MiKTeX* или *TinyTeX*.

Annex I: Подавление предыдущих сигналов

Введение

В данном приложении мы предлагаем метод подавления как часть алгоритма `eears`, используемого в пакете `epitweetr` и описанного выше.

Предположим, $\mathbf{y}$ означает вектор исторических значений длиной в n. Часть вычисления интервала прогнозирования в момент времени 0 – это расчет среднего и стандартного отклонения этих исторических значений, т.е.

$$\overline{y}_0 = \frac{1}{n}\sum_{t=-n}^{-1} y_t \quad \text{и} \quad s_0^2 = \frac{1}{n-1}\sum_{t=-n}^{-1}(y_t - \overline{y}_0)^2$$

Верхний предел одностороннего интервала прогнозирования $(1-\alpha)\times 100\%$ 100% для наблюдения y_0 в рамках модели $y_t \overset{\text{iid}}{\sim} N(\mu, \sigma^2), t = -n, \ldots, 0$ рассчитывается:

$$U_0 = \overline{y}_0 + t_{1-\alpha}(n-1) \times s_0 \times \sqrt{1+\frac{1}{n}},$$

где $t_{1-\alpha}(n-1)$ значит $1-\alpha$ квантиль t-распределения с $n-1$ степеней свободы. Данный расчет порога соответствует статистически верному расчету порога (*Allévius and Höhle* 2017).

Необходимым расширением данного алгоритма является обработка предыдущих сигналов в исторических значениях. Эта проблема уже решалась в рамках квазипуассоновской модели Фаррингтона и др. (1996), сначала применив ОЛМ, а затем повторно применив ОЛМ с весовыми значениями, основанными на остатках

Анскомба. Мы следуем этой общей идее, однако адаптируем ее к отклику по Гауссу, используемому в алгоритме *EARS*, и к соответствующим остаткам линейной модели.

EARS как линейная модель

Мы наблюдаем, что упомянутый выше расчет μ и σ^2 через $\overline{y}_0$ и s_0^2 в момент времени 0 может быть введен в модель линейной регрессии, т.е. для $i = 1, \dots, n$ мы моделируем

$$y_i = \mu + \epsilon_i, \quad \text{где} \quad \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Учтите, что для совместимости со стандартным толкованием теории линейных моделей мы проиндексировали значения y, чтобы y_{-n} соответствовала y_1 и y_{-1} соответствовала y_n. В терминах матрицы допустим, что $\mathbf{y} = (y_1, \dots, y_n)'$ и для модели одного перехвата матрица эксперимента $\mathrm{X} = (1, \dots, 1)'$, которая имеет ранг $k = 1$. Так из стандартной теории линейной регрессии следует:

$$\hat{\mu} \quad = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \,,$$

что соответствует $\overline{y}_0$. Далее предположим, что необработанные остатки определены как $e_i = y_i - \hat{\mu}$ для $i = 1, \dots, n$ и обозначаются $\mathbf{e} = (e_1, \dots, e_n)'$ соответствующим вектором остатков. Следовательно,

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{P}\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{y}$$

где $\mathbf{P} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ – это так называемая проекционная матрица, известная из линейной модели. С помощью этого описания можно записать прогнозируемую величину, σ^2 как описано у *Chatterjee and Hadi* (1988):

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{P})\mathbf{y}}{n-k} = \frac{1}{n-1}\sum_{t=-7}^{1}(y_t - \hat{\mu})^2,$$

что соответствует использованному выше выражению для s_0^2.

Подавление

Мы рассчитываем так называемые **внешние стьюдентизированные остатки** (*Chatterjee and Hadi* 1988)

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-p_{ii}}}, \quad i = 1, \dots, n,$$

где p_{ii} – i-й диагональный элемент проекционной матрицы $\mathbf{P}$ из соответствующей линейной модели, использованной выше. Далее,

$$\hat{\sigma}_{(i)}^2 = \frac{\mathbf{y}_{(i)}{}'(\mathbf{I} - \mathbf{P}_{(i)})\mathbf{y}_{(i)}}{n-k-1}$$

это оценка отклонения, полученная из линейной регрессии, из которой извлечено i-е наблюдение. Теория линейного моделирования (*Chatterjee and Hadi* 1988) заявляет, что

$$r_i^* \overset{\text{идентично}}{\sim} t(n-k-1).$$

Учтите, что остатки одинаково распределены только потому, что они не являются независимыми (см. раздел 4.2.1. *Chatterjee and Hadi* (1988)). Тем не менее, указанная форма распределения позволяет оценить каждое историческое значение, если оно может считаться выбросом. С этой целью мы определяем $r_{\text{порог}}$ как $1-\alpha_{\text{выброс}}$ квантиль t-распределения с $n-k-1$ степеней свободы. Историческое значение выброса (одним возможным объяснением которого является то, что оно происходит из реального увеличения количества твитов, напр., в ситуации выброса), если $r_i^* > r_{\text{порог}}$ Мы будем использовать это, чтобы сформулировать схему весового значения для этих исторических значений:

Подавление-выбросы:

$$w_i^{(\text{dw})} = \begin{cases} 1 & \text{если } r_i^* < r_{\text{порог}} \\ \left(\frac{r_{\text{порог}}}{(r_i^*)}\right)^k & \text{в противном случае} \end{cases}$$
$$= \text{мин.}\left\{1, \left(\frac{r_{\text{порог}}}{r_i^*}\right)^k\right\},$$

где параметр затухания $k > 0$ известен. В оригинальном алгоритме *Farrington et al.* (1996) был использован $k = 2$. Кроме того, использовалось пороговое значение 1. В более поздней работе *Noufaily et al.* (2013) было рекомендовано пороговое значение 2,58. Примечание: оба значения относятся к стандартизированным остаткам Анскомба, которые следуют стандартному нормальному распределению. Если мы берем соответствующие квантили для t-распределения с 6 степенями свободы, значения будут 1,09 и 3,72. Учитывайте также, что компонент $(r_{\text{порог}}/r_i^*)^k$ несколько адаптирован *Farrington et al.* (1996), который вместо него использует $1/(r_i^*)^2$. Преимущество нашего подхода в том, что оно обеспечивает успешную обработку значений в районе порога, если порог не равен 1. Стоит рассмотреть степень выше 2, чтобы обеспечить еще большее подавление для грубых выбросов. Текущее значение параметра затухания по умолчанию в `epitweetr` – 4.

Итак, как в работе *Farrington et al.* (1996), мы нормализуем весовые значения так, чтобы они давали сумму n так

$$w_i^* = n \times \frac{w_i}{\sum_{i=1}^n w_i}$$

а затем снова подгоняем линейную модель с этими весовыми значениями. С этой целью определим весовую матрицу как $\mathbf{W} = \text{diag}(w_1^*, \ldots, w_n^*)$ Мы можем далее использовать взвешенный метод наименьших квадратов, чтобы определить

$$\hat{\mu}_W \quad = (\mathrm{X}'\mathbf{W}\mathbf{X})^{-1}\mathrm{X}'\mathbf{W}\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} w_i^* \, y_i,$$

где есть второй знак равенства, потому что $(\mathrm{X}'\mathbf{W}\mathbf{X}) = \sum_{i=1}^{n} w_i = n$ и $\mathrm{X}'\mathbf{W}\mathbf{y} = \sum_{i=1}^{n} w_i^* \, y_i$. Далее,

$$s_W^2 = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{P}_W)\mathbf{y}}{n-k} = \frac{\sum_{i=1}^{n} w_i^* \, (y_i - \mu_W)^2}{n-1},$$

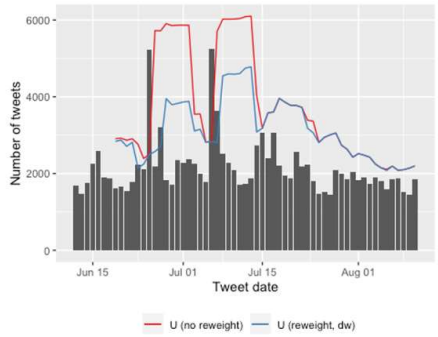
где $P_W = \mathrm{X}(\mathrm{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}\mathbf{W}$ – это проекционная матрица для взвешенных наименьших квадратов.

Так, процедура подавления работает с μ_W и s_W^2 вместо $\overline{y}_0$ и s_0^2, соответственно, при вычислении верхнего предела U_0 с использованием указанной выше формулы.

Пример подхода подавления при использовании данных об Эболе

Рис. 5 ниже показывает верхний предел порога обнаружения сигнала для данных `epitweetr` по Эболе, как оригинальный (красный) без подавления, так и верхний порог с подавлением (синий) после учета предыдущих сигналов в исторических значениях. Учтите, что верхний порог с подавлением обнаруживает три дополнительных сигнала по сравнению с оригинальным порогом.

Рис. 5: Верхний предел как с, так и без подавления для данных *epitweetr* по Эболе



Annex II: Устранение неполадок и советы

Данное приложение содержит несколько советов и общих решений ошибок или проблем, с которым могут столкнуться пользователи `epitweetr`, включая объяснение проверок на странице устранения неполадок.

Страница *troubleshoot*

После проведения диагностики на странице устранения неполадок вы можете увидеть проверки и статус следующих аспектов:

- ***scheduler**:* пакет *R* `taskscheduleR` установлен. только для устройств *Windows*

- ***twitter_auth****:* Токен *Twitter* был создан после аутентификации через учетную запись *Twitter* или в приложении разработчика *Twitter*
- ***search_running***: задача по поиску работает
- ***tweets****:* твиты были собраны
- ***os64****: R* – 64 бита
- ***java****: Java* установлена и доступна для *epitweetr*
- ***java64****: Java* 64 бита установлена и доступна для *epitweetr*
- ***java_version***: установленная версия Java совместима с `epitweetr`
- ***winmsvc****:* Пакет *Microsoft Visual C++ 2010 SP1 Redistributable Package* установлен. Только для устройств *Windows*
- ***detect_activation***: цикл обнаружения активирован
- ***detection_running***: задача по обнаружению работает
- ***winutils****: winutils* установлен Если условие ложно, его можно скачать, запустив задачу по обновлению зависимостей. Только для устройств *Windows*
- **java_deps**: установлены зависимости *java*
- ***move_from_temp****:* `epitweetr` может атомарно перемещать файлы из временной папки в директорию данных
- ***geonames****:* База данных *Geonames.org* скачана и проиндексирована
- ***languages****:* языковые векторы скачаны и проиндексированы
- ***geotag***: задача геотега успешно запущена
- ***aggregate***: задача по агрегированию успешно запущена
- ***alerts***: оповещения были созданы
- ***pandoc****: Pandoc* установлен и доступен для *epitweetr.* Необходимо для создания *PDF*.
- ***tex***: дистрибутив *tex* установлен и доступен для *epitweetr.* Необходимо для создания *PDF*.

Управление поиском и циклами обнаружения (*Windows*)

После включения канала поиска и обнаружения из страницы конфигурации `epitweetr` (*Windows*), будут созданы две задачи в планировщике задач и будут предложены два окна терминала. Учтите, что если вы вышли из сеанса на компьютере/выключили его или закрыли окна терминала, каналы поиска и обнаружения остановятся.

Если вы снова активируете эти задачи на странице конфигурации `epitweetr`, система перепишет задачи, созданные в планировщике задач. Наоборот, после первой успешной активации этих задач из `epitweetr`, вы сможете легко управлять ими из планировщика задач. Вы можете остановить эти задачи, завершив и отключив их в планировщике задач, и возобновить эти задачи, включив и запустив их в планировщике задач.

В планировщике задач вы можете установить, что задачи поиска и обнаружения будут «*run whether the user is logged on or not*», чтобы избежать остановки этих задач при выходе из сеанса или перезагрузке компьютера.

Управление поиском и циклами обнаружения (*Linux* и *Mac*)

Так как канал поиска и обнаружения в *Linux* или *Mac* требуют запуска вручную, если вы вышли из сеанса на компьютере/выключили его или окна терминала закрыты, поиск и канал обнаружения остановятся. Помните о необходимость следовать шагам в разделе *Setting up tweet collection and the alert detection loop*, чтобы возобновить эти задачи.

Работа поиска и канала обнаружения

«Cannot execute task #####: the task is already running»

Цикл обнаружения поиска создает два файла, которые содержат *ID* процессов, находящихся в папке `epitwitter`: *search.PID* и *detect.PID*. Эта ошибка появляется, если `epitweetr` находит другой процесс *R* работающий под таким же *ID*. Чтобы исправить эту ошибку, вы должны сначала проверить, не работает ли цикл поиска/обнаружения в другой сессии *R*. В этом случае вам не стоит запускать задачу, так как `epitweetr` поддерживает только одну копию задачи, запущенной в одной и той же папке данных. Если работающий процесс не связан с задачей, вы можете вручную удалить файл *PID* и перезапустить его.

Ошибка при попытке агрегирования файлов

Это может произойти по двум причинам: – Недостаточно места на диске для временных файлов. Процесс агрегирования создает временные файлы перед сохранением их в соответствующей папке *epitweetr*. В таком случае измените переменную среды своей учетной записи для *TMP* и *TEMP* на другое место с большим объемом. – Если ошибка появляется при создании конкретного файла, то может существовать поврежденный серийный файл для этой даты. Удалите *'country_counts'*, *'geolocated'* и *'topwords'* для этой даты и перезапустите задачу вручную, нажав на соответствующую кнопку на странице конфигурации.

Изменение пользователя для аутентификации *Twitter* при использовании учетной записи *Twitter*

1. Завершите и отключите цикл/задачу обнаружения в планировщике задач (*Windows*) или закройте R/окно терминала с помощью цикла/задачи поиска (*Linux* и *Mac*)

2. Найдите файл «*.rtweet_token*» в скрытых файлах. Он обычно находится в папке Документы.

3. Удалите этот файл.

4. Нажмите на «*Update properties*» на странице конфигурации `epitweetr`.

5. Включите и запустите цикл/задачу обнаружения в планировщике задач (*Windows*) или запустите R/окно терминала с помощью цикла/задачи поиска (*Linux* и *Mac*). Более детальная информация доступна в разделе «*Setting up tweet collection and the alert detection loop*»

Скачивание *GeoNames* и/или языков

«The specified size exceeds the maximum representable size. Error: Could not create the Java Virtual Machine»

Если эта ошибка появляется при запуске *GeoNames*, это значит, что на компьютере установлена *Java 32bits*. Вам необходима *Java 64bits*. Сделайте ее доступной для *epitwitter*, установив переменную среды «*JAVA_HOME*» или настроив правильный двоичный файл *java* в системном *PATH*.

«Max number of retried reached failed while processing languages. Error in get_geolocated_period(dataset): To aggregate, or calculate alerts geolocation must have been successfully executed, but no geolocation files were found»

Если вы видите эту ошибку на странице конфигурации, это значит, что `epitweetr` не может найти геотеги собранных твитов. Крайне рекомендуется перезапустить *GeoNames* и языки, так как они могли быть скачаны с ошибками. При запуске этой задачи необходимо убедиться, что вы не вышли из сеанса на компьютере/выключили его или он не перешел в спящий режим.

«*Launch slots*» на странице конфигурации показывают *NA* вместо временных слотов

При первой установке и запуске `epitweetr`, задача поиска геотега канала обнаружения должна быть запущена хотя бы раз, чтобы увидеть временные слоты в «Запуске слотов» («*Launch slots*») на странице конфигурации.

Скачивание *PDF* панели управления

«Error in: LaTeX failed to compile C:\Users\name~1\...\file######.tex.»

Эта ошибка появляется в *Windows* при нажатии на «*PDF*» в панели управления, однако *PDF* не сохраняется. Причиной этого может быть слишком длинный путь к переменным среды *TEMP* и *TMP*, при котором *Windows* сокращает путь и *epitweetr* не может найти новый путь. Для исправления ошибки выполните следующие шаги:

1. Откройте «переменную среды для вашей учетной записи»
2. Измените путь для *TEMP* и *TMP* на более короткий (напр., «*C:\Temp*»). Такой же путь должен использоваться для обеих переменных среды.
3. Выйдете и войдите в сеанс
4. Вы можете скачивать и сохранять *PDF* из панели управления

«Error: pandoc document conversion failed with error 6»

1. Скачивание скрипта (https://raw.githubusercontent.com/jgm/pandoc/master/macos/uninstall-pandoc.pl)

2. Удалите *pandoc* (https://pandoc.org/installing.html, запустив перл (*perl*) uninstall-pandoc.pl

Разные общее количество выводимых данных панели управления

При расчете общего количества твитов в панели управления приложения *Shiny app* или в скачиваемых данных вы можете получить различия в общем количестве твитов между тремя выводимыми данными. Это может произойти по следующим причинам:

1. *World (all)* против *World (geolocated)*

    - Опция по умолчанию для регионов в *World (all)*значит, что в линию тренда включаются также твиты без геолокации, однако на картах и в числе самых частотных слов визуализируются только твиты с геолокацией, поэтому общее количество твитов может различаться между выводимыми данными при выборе *World (all)* или пустого значения по умолчанию.
2. Анализ по конкретной стране

    - При выборе только одной страны в фильтрах линия тренда будет показывать все твиты для этой страны, но на карте будут показаны твиты на субнациональном уровне. Может быть, что геолокация некоторых твитов была привязана к определенной стране без дополнительных данных на субнациональном уровне. Эти твиты будут показаны на итоговой линии тренда, но не в кругах на субнациональном уровне карты.
3. Самые частотные слова

    - В отличие от других выводимых данных на панели инструментов, число самых частотных слов всегда основано на местоположении твита независимо от фильтра (по причине объема памяти). Поэтому, если в фильтре местоположения выбрано местоположение пользователя или оба местоположения, эта цифра может иметь иное общее значение, чем два других вида выводимых данных.

Получение только оповещений в режиме реального времени

Это связано с пользователями, которые выбрали темы и/или регионы для получения оповещений в режиме реального времени или выбрали темы и/или регионы для получения оповещений по расписанию. Если в таких случаях вы получаете только оповещения в режиме реального времени со всеми темами и регионами, вероятно, никакие временные слоты не были включены в файл

подписчиков со страницы конфигурации. Эти временные слоты используются для планированных оповещений, и если никакие временные слоты не были включены в файл, то оповещения по всем темам и регионам отправляются как оповещения в режиме реального времени.

Неполучение оповещений по эл. почте

Если вы не получаете оповещения по эл. почте и видите ошибку в `epitweetr` обозначающий отклоненный вход в учетную запись, это значит, что `epitweetr` не может войти в учетную запись эл. почты на странице конфигурации. Некоторые причины этого:

- Сервер или порт, указанные на странице конфигурации, неверны
- Попытка `epitweetr` войти в учетную запись эл. почты блокируется сервером. Это может случиться с учетными записями эл. почты некоторых организаций. В таком случае, свяжитесь с IT-департаментом своей организации.
- При использовании учетной записи вам необходимо разрешить использование менее безопасных приложений в настройках своей учетной записи.

Ссылки

Allévius, Benjamin, and Michael Höhle. 2017. “Prospective Detection of Outbreaks.” *arXiv:1711.08960 [Stat]*, ноябрь. https://arxiv.org/abs/1711.08960.

Chatterjee, Samprit, and Ali S. Hadi. 1988. *Sensitivity Analysis in Linear Regression*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Farrington, C. P., N. J. Andrews, A. D. Beale, and M. A. Catchpole. 1996. “A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease.” *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159 (3): 547. https://doi.org/10.2307/2983331.

Fricker, Ronald D., Benjamin L. Hegler, and David A. Dunfee. 2008. “Comparing Syndromic Surveillance Detection Methods: EARS’ Versus a CUSUM-Based Methodology.” *Statistics in Medicine* 27 (17): 3407–29. https://doi.org/10.1002/sim.3197.

Noufaily, Angela, Doyo Enki, Paddy Farrington, Paul Garthwaite, Nick Andrews, and Andre Charlett. 2013. “An Improved Algorithm for Outbreak Detection in Multiple Surveillance Systems.” *Online Journal of Public Health Informatics* 5 (1): e148. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3692796/.

Salmon, Maëlle, Dirk Schumacher, and Michael Höhle. 2016. “Monitoring Count Time Series in *R* : Aberration Detection in Public Health Surveillance.” *Journal of Statistical Software* 70 (10). https://doi.org/10.18637/jss.v070.i10.