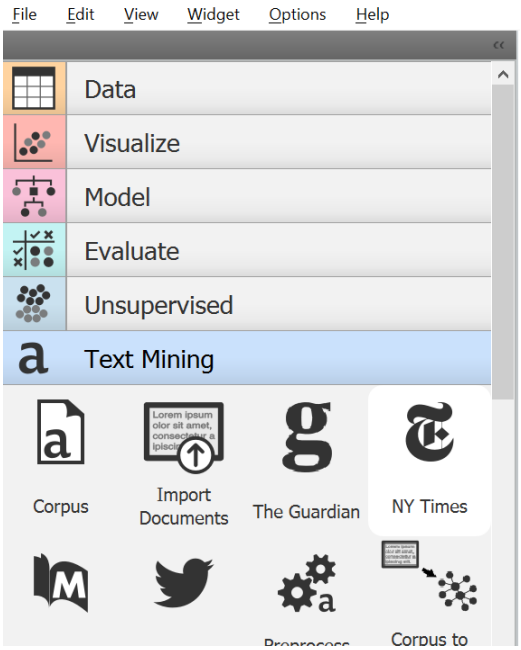
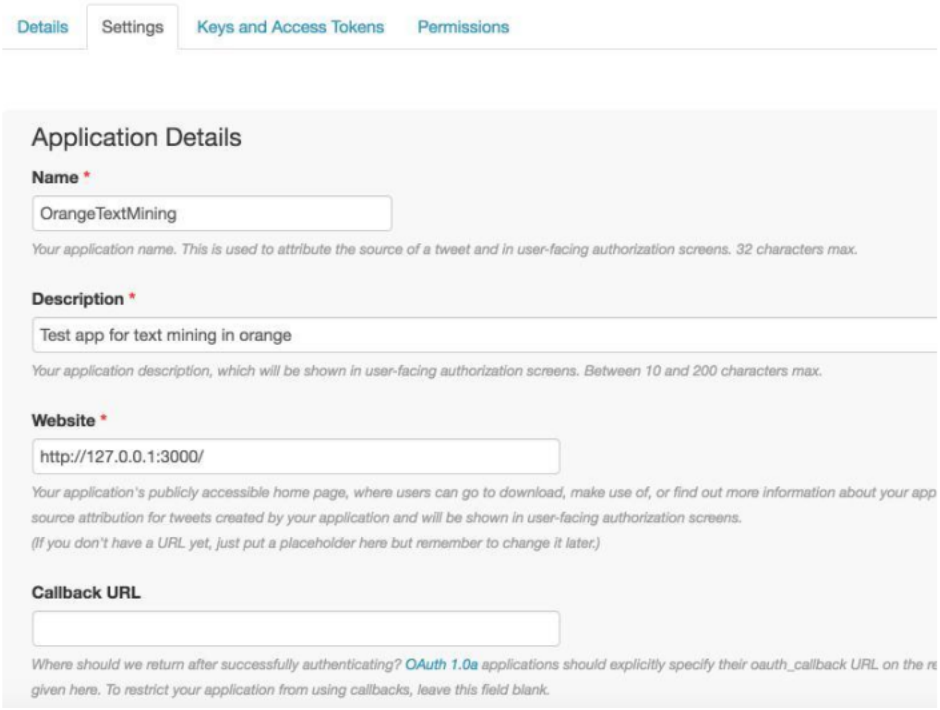
Лабораторная работа №3
ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТА. WEB MINING.
ТОНАЛЬНЫЙ АНАЛИЗ ТЕКСТА В ORANGE

Цель и задача работы Изучить основные методы Web Mining и Text Mining с использованием приложения «Orange Data Mining». Используя разные методы текстового анализа, проанализировать реакцию пользователей социальной сети Twitter на различные события (согласно полученному заданию). Осуществить поиск закономерностей в текстовых документах.

Построения модели Text Mining в Orange Для получения доступа к инструментам текстового анализа в **Orange** необходимо подключить соответствующее дополнение **Orange3-Text**. Для этого откройте окно управления дополнениями (**Options** > **Add-ons**). Выберите дополнение **Orange3-Text.** После этого появится следующая вкладка



Теперь необходимо создать приложение Twitter для дальнейшей работы. Для этого нужно зарегистрироваться в социальной сети Twitter и пройти на страницу https://apps.twitter.com/ (или https://developer.twitter.com/en/apps) Далее нажать кнопку **Create New App** и ввести короткие сведения о приложении. Пример показан на рис.



В поле Website можно указать любой адрес, ссылающийся на компьютер пользователя. Например, http://127.0.0.1:xxxx/, где xxxx – произвольный номер порта. Значение может быть любым, так как всё равно дальше использоваться не будет. После этого на вкладке **Keys and Access Tokens** будут доступны данные для доступа к API Twitter. Нам понадобятся API key и API secret. Их необходимо скопировать и ввести в окне настройки виджета **Twitter** (кнопка **Twitter API key**). Далее вводим условия поиска сообщений (по одному на каждой строке) в поле **Query word list**. В качестве языка указываем Russian. Максимальное количество сообщений (**Max tweets**) желательно выбрать в диапазоне от 200 до 2000.

После этого можно нажать кнопку **Search** и посмотреть, сколько сообщений было найдено по предложенному поисковому запросу.

Подготовка сообщений к анализу (виджет Preprocess Text) Для подготовки данных к дальнейшему анализу необходимо очистить их и привести к нужному виду. Для этого воспользуемся виджетом **Preprocess Text**.

Откроем окно настроек виджета **Preprocess Text** (рис.). Осуществим преобразование всех слов к нижнему регистру (Lowercase) и удалим из сообщений адреса веб-страниц (Remove urls). В качестве режима лексического анализа (Tokenization) можно выбрать две опции **Regexp** (введенное регулярное выражение будет использовано для выделения токенов) или **Tweet** (предопределённая модель, которая сохранит хэштеги, смайлы и другие специальные символы). Для тонального анализа подходит режим **Tweet**.

В группе **Filtering** можно задать различные варианты фильтрации и очистки сообщений. Так, можно составить список **Stopwords** (слова, которые

будут удалены из сообщений) в виде txt-файла. Можно задать список разрешенных слов (**Lexicon**) также в виде txt-файла. Опция **Regexp** позволяет удалить служебные символы (кавычки, числа, пунктуацию и т.д.). Обязательно используйте регулярные выражения для очистки сообщений.

Поиск топиков (виджеты Bag of Words, Topic Modeling и Word Cloud) Теперь есть данные, готовые к обработке. Для начала выделим наиболее частые лексические повторения в полученных сообщениях. Для этого необходимо добавить три новых виджета:

- bag of Words (позволяет выделить наиболее употребляемые, характерные для многих сообщений слова);
- topic Modeling (находит наиболее характерные для всех сообщений наборы слов);
- word Cloud (позволяет отобразить слова в виде облака, где самые характерные слова располагаются максимально близко к центру облака и имеют наибольший размер, а наименее характерные – расположены далеко от центра и имеют минимальный размер).

Теперь можно открыть окно настроек виджета **Topic Modeling** и посмотреть найденные наборы слов. В случае, если в наборах слов встречаются знаки препинания или служебные символы, их необходимо исключить с помощью изменения регулярного выражения для фильтрации виджета **Preprocess Text**.

Этих данных уже достаточно для некоторого анализа отношения пользователей Twitter к предмету исследования. Для лучшей визуальной картины можно представить полученные результаты в виде облака слов (**Word Cloud**).

Проанализировав полученное облако слов, можно улучшить результаты анализа, добавив список **Stopwords**.

Тональный анализ (виджет Tweet Profiler) Для определения тональной окраски сообщений воспользуемся виджетом **Tweet Profiler** (он основан на анализе иконок эмоций или, проще говоря, смайлов). Выходные данные представим в виде точечной диаграммы (**Scatter Plot**).

Остановимся подробнее на настройках виджета **Tweet Profile** (рис.). Для нормальной работы данного виджета необходимо получить токен доступа, для этого просто нужно нажать кнопку **Get Token.** Кроме того, виджет предлагает выбрать шкалу тональности (**Emotions**). Для выбора доступны шкалы Пола Экмана, Роберта Плутчика и современные Профили Настроения (Profile of Mood States или POMS). Попробуйте применить каждую из них и посмотреть на результат. Для визуализации полученного результата воспользуемся

точечной диаграммой. На рис. показан пример диаграммы. Для удобства представления по оси *x* (Axis *x*) выбираем поле Emotion, а по оси *y* (Axis *y*) выбираем поле Date.

Полученная диаграмма является визуальным представлением общей тональной окраски всех полученных сообщений. На основании этой диаграммы можно сделать вывод об общем отношении пользователей к предмету исследования.

Порядок выполнения работы

1. Получить исходные данные (данные о каком-либо событии/фильме/личности и т.д.) с помощью виджета **Twitter**.
2. Осуществить необходимую фильтрацию/очистку данных с помощью виджета **Preprocess Text.**
3. Осуществить поиск топиков с помощью виджетов **Bag of Words** и **Topic Modeling**.
4. Вывести результаты с помощью виджета **Word Cloud**.
5. Осуществить тональный анализ данных с помощью виджета **Tweet Profiler**.
6. Вывести результаты с помощью точечной диаграммы (виджет **Scatter Plot**).
7. Сравнить результаты тонального анализа и поиска топиков. Сделать выводы.
8. Выбрать фильмы/личности/события с противоположной эмоциональной окраской и осуществить тональный анализ.
9. Найти существующие закономерности в твитах, объяснить их, вывести самостоятельно с помощью цепочки виджетов.
10. Подготовить отчёт.