Учреждение образования

«Белорусский государственный университет

информатики и радиоэлектроники»

Кафедра программного обеспечения информационных технологий

Отчет по лабораторной работе №4

по предмету

«Стили и методы Программирования»

Выполнила: Манин Алексей Сергеевич

магистрант кафедры ПОИТ

группа №857041

Проверила: Парамонов Антон Иванович

кандидат технических наук, доцент

Минск 2019

**Задание.**

Подключиться к одному из ресурсов сети Интернет (например, социальной сети). Использовать средства существующих библиотек языка. Загрузить тестовую выборку данных с ресурса (например, связи пользователей или открытые данные пользователей) и сохранить на локальном хранилище. Загрузить данные из файла средствами языка в рабочий процесс скрипта. Провести анализ выборки (например, кластеризация пользователей или статистика по заданным параметрам). Визуализировать результаты анализа внешними библиотеками (например, построение диаграмм или графиков).

**Индивидуальное задание**

В рамках данной лабораторной работы анализируются студийные альбомы группы Death Grips в контексте звучания и лексического состава песен, с использованием данных, предоставляемых стриминговым сервисом Spotify и веб-сайтом Genius – краудсорсинговой библиотекой текстов песен.

**Порядок выполнения**

Ход работы можно условно разделить на три этапа.

**Первый этап: извлечение и подготовка данных**

Первый этап состоит в подключении к сервисам Spotify и Genius и последующем получении наборов данных с каждого из них. С этой целью используется R-пакет spotifyr, который включает в себя в том числе пакет genius для получения данных с соответствующего сервиса.

Каждый из ресурсов требует регистрации пользовательского приложения для получения токенов доступа.

После необходимых установок, непосредственно извлечение данных Spotify выполняется следующим образом:

dg <- get\_artist\_audio\_features('death grips')

С помощью приведенной команды, очевидно, получается набор данных о песнях избранных исполнителей. Полученный dataframe включает в себя различные значения, в том числе название песни, название альбома, продолжительность, музыкальный размер, а также ряд звуковых характеристик, сгенерированный Spotify для всех доступных на сервисе песен, среди которых имеются энергичность (англ. energy), танцевальность (англ. danceability), валентность (англ. valence) и другие.

Полученный набор данных фильтруем таким образом, чтобы оставить только песни со студийных альбомов группы.

Данные с ресурса Genius извлекаются мануально следующим образом:

genius\_get\_artists <- function(artist\_name, n\_results = 10) {

baseURL <- 'https://api.genius.com/search?q='

requestURL <- paste0(baseURL, gsub(' ', '%20', artist\_name),

'&per\_page=', n\_results,

'&access\_token=', token)

res <- GET(requestURL) %>% content %>% .$response %>% .$hits

map\_df(1:length(res), function(x) {

tmp <- res[[x]]$result$primary\_artist

list(

artist\_id = tmp$id,

artist\_name = tmp$name

)

}) %>% unique

}

genius\_artists <- genius\_get\_artists('death grips')

С помощью приведенного кода получаем идентификатор артиста на сервисе. Далее для выбранного артиста получается список песен, который после базовой обработки (например, корректировка названий песен, различающихся на обоих сервисах) фильтруется на основе данных, полученных на Spotify (аналогично оператору SQL LEFT JOIN). Наконец, на основе отфильтрованного списка песен, для каждой из них с Genius извлекается текст песен. Из текстов удаляются стоп-слова, например «oh», «yeah», «hey» и так далее. Удаляются специфичные для Genius символы разметки и обозначения куплетов, припевов.

В результате описанных манипуляций, имеем данные, готовые для анализа.

**Второй этап: анализ звучания**

На втором этапе на основе данных о характере звучания песен, полученных со Spotify, анализируются альбомы артиста и песни в отдельности.

Прежде чем приступить к описанию выполняемых вычислений, приведем определение музыкальной валентности с официального сайта Spotify (приводится примерный перевод):

Валентность – это мера от 0,0 до 1,0, описывающая музыкальную позитивность песни. Песни с высокой валентностью звучат более позитивно (счастливые, веселые, эйфорические), в то время как песни с низкой валентностью звучат более негативно (грустные, депрессивные, злые).

Так, в рамках анализа звучания подсчитывается средняя валентность, строятся графики Joy (или Ridgeline Plot) для каждого из альбомов. Как видно из полученных графиков, звучание Death Grips на каждом из альбомов тяготеет к негативному. Средние значения валентности для альбомов простираются от 0.3 до 0.45, где наиболее депрессивный альбом – The Powers That B, а наиболее позитивный (или скорее наименее негативный) – The Money Store. Альбом The Money Store – первый полноценный альбом группы, – является единственным в дискографии группы, выпущенным на мэйджор-лейбле EPIC Records (входит в состав Sony Music Entertainment, список исполнителей лейбла включает таких артистов, как AC/DC, Judas Priest и Селин Дион), остальные же альбомы были выпущены самостоятельно и, соответственно, без каких-либо контрактных обязательств перед издателем.

С точки зрения валентности, печаль и гнев – это одно и то же. Набор данных Spotify кроме всего прочего содержит индекс энергичности. Предположим, что печальные песни звучат менее энергично, тогда как агрессивные или гневные – более энергично. Подсчитываем среднюю энергичность песен на каждом из альбомов. Как видно из полученных результатов, значения средней энергичности для альбомов принадлежат промежутку от 0.8 до 0.93. Очевидно, музыка группы Death Grips гневная, агрессивная.

Совместим значения валентности и энергичности в единый параметр «anger index» (индекс злости). Данный параметр будем высчитывать следующим образом:

.

Валентность отнимается от единицы для того, чтобы большие средние значения были более негативными. Таким образом, самыми «злыми» песнями будут высоко-энергичные и низко-валентные.

В заключение, введем коэффициент танцевальности в анализ. Как видно из полученных результатов, альбом The Money Store является наиболее танцевальным. Данный результат коррелирует с информацией, приведенной ранее: первый альбом группы был выпущен на крупном лейбле популярной музыки и является наиболее доступным для прослушивания с точки зрения звучания (более мэйнстримовым).

**Третий этап: лексический анализ**

Тексты песен, полученные от Genius, разбиваются на отдельные слова. Тексты песен предварительно были очищенные от стоп-слов на первом этапе выполнения лабораторной работы, как указано ранее. Для такой фильтрации использовался dataframe STOP\_WORDS из R-пакета tidytext.

Приступим к анализу. Первым делом, построим облако слов, наиболее часто встречающихся в текстах группы. Как видно, в текстах присутствует много (очень много) обсценной лексики. Наиболее популярные слова преимущественно имеют негативную направленность.

Текстам свойственна повторяемость. Такие выводы можно сделать на основе графика лексического разнообразия альбомов. Как видно на графике, разнообразие текстов оценивается в среднем в промежутке от 0.2 до 0.5.

Дополнительно, тексты тестируются на «эмоциональный окрас» на базе специализированных словарей (англ. sentiment lexicons) AFINN, bing и nrc.

Словарь AFINN присваивает каждому слову значения от -5 до +5 баллов в зависимости от того, насколько негативное или позитивное значение это слово несет. Например, слово «шедевр» имеет оценку +4, тогда как «мука» – -4. Нейтральные слова имеют оценку 0. В работе словарь AFINN используется для создания гистограммы сумм баллов всех слов в альбоме. Каждое слово учитывается только один раз. Как видно, в общем лексически альбомы резко негативные.

Словарь bing описывает каждое слово в категориях «позитивное» или «негативное». В работе данные bing используются для построения «пирамиды населенности». Значения слева от оси – это число негативных слов в песне, справа от оси – позитивных. Анализ на основе bing произведен для альбома The Money Store. Выделяется песня «Bitch Please», в которой на основе словаря bing не обнаружено ни одного позитивного слова.

Словари AFINN и bing весьма ограничены, поскольку в них слова рассматриваются в «бинарном» представлении «положительное» или «отрицательное». Словарь nrc классифицирует слова на базе восьми категорий: радость, предвкушение, доверие, удивление, грусть, злость, отвращение и страх. Проанализировав тексты песен с использованием nrc, построим на основе полученных данных радиальную диаграмму, чтобы продемонстрировать, каким образом каждый из альбомов располагаются относительно друг друга в контексте этих восьми «эмоций». Каждая вершина представляет собой процент слов, относящихся к конкретной категории. Как видно из полученного графика, лексически тексты песен также тяготеют к негативным эмоциям. На основе полученных данных трудно проследить динамику развития «негатива» от альбома к альбому, так как все они в равной степени негативно-ориентированы. Ярко выражена тема страха на первых двух альбомах.

**Выводы**

Группа Death Grips с 2011 года выпускает инновационную, интересную и, что особенно важно в рамках данной работы, грубую, агрессивную, резкую музыку. Абразивность звука Death Grips – выражающаяся в форме использования уникальных сэмплов, искаженных синтезаторов и почти неразборчивых криков солиста, - часто приводит к тому, что слушатели, услышав музыку группы впервые, отказываются от последующего их прослушивания.

Полученные в результате работы данные подтверждают тот факт, что музыка группы является тяжелой как с точки зрения звучания, так и с точки зрения смысла, заложенного в ней, и, как следствие, сложной для восприятия (как минимум психологически).

Несмотря на то, что у многих людей существует в каком-то смысле «травмирующее» представление о музыке Death Grips, творчество группы обрело культовое следование фанатов по всему миру, а музыкальные критики называют их одними из самых влиятельных артистов в современной музыке. Этот факт побуждает тех же слушателей вернуться к музыке Death Grips, и с каждым последующим прослушиванием слушатели получают все больше удовольствия от данного опыта.

В работе использовался ряд R-пакетов для визуализации данных, таких как ggplot2, yarrr (функция pirateplot), wordcloud и radarchart. Графики, полученные с их использованием, можно найти в приложенном к отчету архиве. Кроме того, для создания и форматирования html-таблиц с различными списками для вывода, использовалась библиотека kableExtra. В работе активно использовалась библиотека dplyr (вместе с ggplot2 и другими полезными пакетами входящая в состав супер-пакета tidyverse), значительным образом упрощающая работу с данными в R в первую очередь, на мой взгляд, благодаря оператору последовательного выполнения %>%. Для лексического анализа использовался пакет tidytext, включающий в себя дэйта-фрэймы специализированных словарей (в частности AFINN, bing и nrc) для словесного анализа. Каждый промежуточный итог лабораторной работы был визуализирован в том или ином виде и стилизован с использованием внутренних средств каждого из пакетов, а также с использованием кастомных цветовых схем пакета RColorBrewer.

Скрипт R приложен к отчету.