Ход работы

Описание источника данных

Данные были собраны из социальной сети Twitter, которая позволяет пользователем делать публикации длиной до 280 символов. Также пользователи могут добавлять в свой профиль личную информацию.

Описание полученных из источника данных

В результате работы были собраны данные о пользователях социальной сети Twitter, которые включают в себя информацию и описание профилей, а также одну из последних публикаций и информацию о ней для каждого из пользователей.

Таблица 1 – Статические характеристики массива данных

Число элементов	16597
Общий размер набора	33.1 Мб
Средний размер элемента	2 Кб
Среднее число слов в элементе	145
Количество полей для каждого элемента	72

Описание обобщенной формализованной структуры полученных данных

Данные были собраны и сохранены в один файл (my_data.JSON) в текстовом формате JSON, где каждый элемент выборки представляется в виде отдельного объекта.

Структура полученных данных:

```
object
    '{' members '}'
members
    member
    member ',' members
member
    string ':' element
string
    ' " ' characters ' " '
element
    value
value
    object
    array
    string
    number
    "true"
    "false"
    "null"
array
    '[' elements ']'
elements
    element
    element ',' elements
```

Этап очистки данных

Правила:

- если встретили подстроку, заданную шаблоном <.*?>, в полях text и description, то заменяем найденную подстроку на пустую (удаляем «смайлики»);
- если встретили подстроку, заданную шаблоном @.*?\\s, в полях text и description, то заменяем найденную подстроку на пустую (удаляем ссылку на другого пользователя);
- если встречаем символ из подстроки [][!#$%&()*,:;<=>@^_`’|~{}+-] в полях text и description, то заменяем его на пустой (убираем ненужные для целей обработки символы);
- если встретили подстроку, заданную шаблонами https.*?\\s и https.*?&, в полях text и description, то заменяем найденную подстроку на пустую (убираем ссылки);
- если встретили подстроку, заданную шаблоном \n, в полях text и description, то заменяем найденную подстроку на пустую (удаляем символ переноса строка);
- если встретили подстроку, заданную шаблоном \\s’.*?\\s, в полях text и description, то заменяем найденную подстроку на пустую (удаляем отдельно стоящие сокращения, которые могли остаться при применение других правил очистки, например, «‘s»);
- если встретили слово из списка стоп-слов в полях text и description, то удаляем его (удаляем предлоги, союзы и т.д., список стоп-слов взят из пакета языка R);
- если встретили подстроку `Twitter` или `for` в поле source, то удаляем эту подстроку (удаляем повторяющиеся термины, которые не несут смысловой нагрузки, в поле каждого элемента).

Этап нормализации данных

Правила:

- если встретили прописную букву в полях text и description, то заменяем ее на строчную (приводим данные внутри полей к единообразному виду).

Этап удаления «шума»

Правила:

- удалить поля user_id, status_id, reply_to_user_id, is_quote, favorite_count, mentions_user_id, mentions_screen_name, geo_coords, coords_coords, bbox_coords, favourites_count, listed_count, reply_to_status_id, quoted_status_id, quoted_text, quoted_created_at, quoted_source, quoted_favorite_count, quoted_retweet_count,

quoted_user_id, quoted_screen_name, quoted_name, quoted_followers_count, quoted_friends_count, quoted_location, quoted_statuses_count, quoted_description, quoted_verified, country_code, place_type, place_name, place_full_name – данные поля не несут смысловой нагрузки для цели аналитического исследования;

- удалить поле created_at – значение данного поля для каждого элемента выборки одинаково (не считая секунд) - 2018-09-18 19:31:59, так как собирались публикации от одного времени;
- удалить поля urls_url, urls_t.co, urls_expanded_url, media_url, media_t.o, media_expanded_url, ext_media_url, ext_media_t.co, ext_media_expanded_url, status_url, profile_background_url, profile_image_url, url, profile_url, profile_banner_url, profile_expanded_url, place_url – данные поля содержат какие-либо ссылки, информация о которых не важна для аналитической цели исследования;
- удалить поля is_retweet, retweet_count, symbols, protected, verified – значение полей одинаково для всех элементов;
- удалить поле lang – в процессе собирались публикации на английском языке, следовательно, это поле для всех элементов одинаково – en.

Оставшиеся поле после этапа «удаление шума» и их описание:

`"screen_name": "CodiVbug"` – ник пользователя в социальной сети, string;

`"text": "@DangeRussWilson @Seahawks @PeteCarroll we still love you all! Use the haters as motivation to be better today then yesterday make them eat their words and pray they find faith in our team!"` – содержание публикации пользователя, string;

`"source": "Twitter for Android"` – описание устройства, с которого был опубликован пост, string;

`"display_text_width": 190` – количество символов в публикации, number;

`"reply_to_screen_name": "DangeRussWilson"` – ник пользователя, которого упомянул в своей публикации текущий пользователь, string;

`"hashtags": [null]` – хэштэги, которые были упомянуты в публикации, number либо null;

`"media_type": [photo]` - тип прикрепленных к публикации приложение, array[string];

`"name": "codi vermeer"` – имя пользователя, string;

`"location": ""` – название места (город, область и т.д.), в котором находился пользователь в момент публикации поста, string;

"description": "" – описание профиля пользователя, string;

"followers_count": 16 – количество пользователей, подписанных на данный аккаунт, number;

"friends_count": 115 – количество пользователей, на которых подписан данный аккаунт, number;

"statuses_count": 180 – количество публикаций у данного пользователя, number;

"account_created_at": "2013-06-03 02:10:05" – дата создания данного аккаунта, формат даты и времени (год-месяц-число часы:минуты:секунды);

"account_lang": "en" – язык данного аккаунта, везде – en, string.

Для этапов очистки, нормализации данных и удаления «шума» использовался язык R для статистической обработки данных. Преобразования производились в среде разработки RStudio для Windows. Все данные из JSON формата были преобразованы в таблицу (CSV). Для удаления «лишних» полей для каждого элемента получившейся таблицы использовались стандартные функции языка R.

Использованные библиотеки:

- twitteR, RCurl, ROAuth, rtweet, httpuv (для сбора данных из Twitter);
- tm (для загрузки словаря стоп-слов для этапа очистки данных);
- jsonlite (для работы с собранными данными в формате JSON).

Анализ результатов очистки данных

Таблица 2 – Результаты очистки данных

Параметр	**Исходный набор**	**Очищенный набор**	**Нормализованный набор**	**Набор без «шума»**
Размер набора - байт	33.1 Мб	32.1 Мб	33.2 Мб	10.2 Мб
Число элементов	16597	16597	16597	16597
Размер элемента (средний)	2 Кб	1.9 Кб	2 Кб	0.6 Кб
Размер элемента (максимальный)	2.37 Кб	2.27 Кб	2.36 Кб	908 байт
Размер элемента (минимальный)	1.8 КБ	1.64 Кб	1.81 Кб	541 байт
Число фактов в элементе (среднее)	158	138	158	74
Число фактов (слов, единиц языка) в элементе (минимальное)	130	106	130	52
Число фактов (слов, единиц языка) в элементе (максимальное)	186	171	186	97

Нормализованный набор данных не отличается по размеру от исходного набора данных, так как проводилась лишь замена прописных букв на строчные. Наилучшие результаты с точки зрения уменьшения размера набора показал этап «удаления шума». Так как в ходе работы на данном этапе были удалены больше половины исходных полей. Этап «очистки» позволил уменьшить размер исходного набора примерно на 1 Мб, а также уменьшить количество фактов в каждом элементе до 138 (примерно).

Перечень сущностей и наборы описывающих их данных

Были определены две сущности, которые описываются собранными данными, а именно пользователь и его публикация. Также были определены компоненты, описывающие каждую из сущностей.

Пользователь:

```
{
	"screen_name": string;
	"name": string;
	"description": string;
	"friends_count": number;
	"followers_count": number;
	"statuses_count": number;
	"account_created_at": string;
	"account_lang": string;
	"country": string;
}
```

Публикация:

```
{
	"text": string;
	"source": string;
	"display_text_width": number;
	"hashtags": array[string];
	"media_type": array[string];
	"location": string;
}
```

Компоненты собранных данных одинаково подробно описывают, как сущность «Пользователь» (9 компонент), так и сущность «Публикация» (6 компонент).

Таблица 3 – Компоненты данных и их тип

Компонент данных	**Тип**	**Способ кодирования (если есть)**
screen_name	Полнотекстовое значение	-
name	Полнотекстовое значение	-
description	Полнотекстовое значение	-
friends_count	Число	-
followers_count	Число	-
statuses_count	Число	-
account_created_at	Дата и время	-
account_lang	Словарное значение	Уровневое кодирование
country	Словарное значение	-
text	Полнотекстовое значение	-

source	Значение из предопределенного списка	-
display_text_width	Число	-
hashtags	Полнотекстовое значение	-
media_type	Значение из предопределенного списка	Уровневое кодирование
location	Полнотекстовое значение	-

Был введен новый тип «дата и время» для компонента данных «account_created_at». Данное поле имеет одинаковую структуру для каждого из элемента и представляет собой формат даты и времени в числовых значениях.

Кодирование

Для поля «account_lang» было выбрано уровневое кодирование. Для того, чтобы учесть семантическую близость значений данного поля, была построена иерархическая классификация. На ее основе возможные значения данного поля были закодированы при помощи кодов, представленных в Приложении А. Так как значение данного поля – это язык, на котором ведется данный аккаунт в Twitter, то было решено разделить значения в зависимости от того, в какой части мира говорят на данном языке. Для возможности добавления нового языка используется кодирование десятичными числами, разделенными точками, где каждое число обозначает соответственно: часть света, регион и язык.

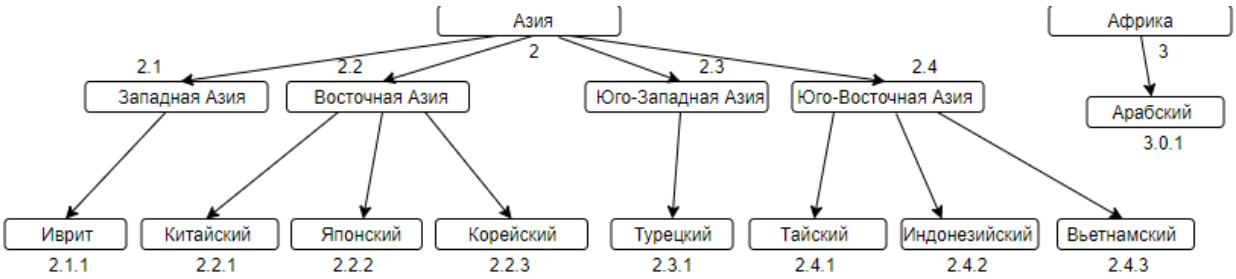

Рисунок 1 – Иерархическая классификация языка

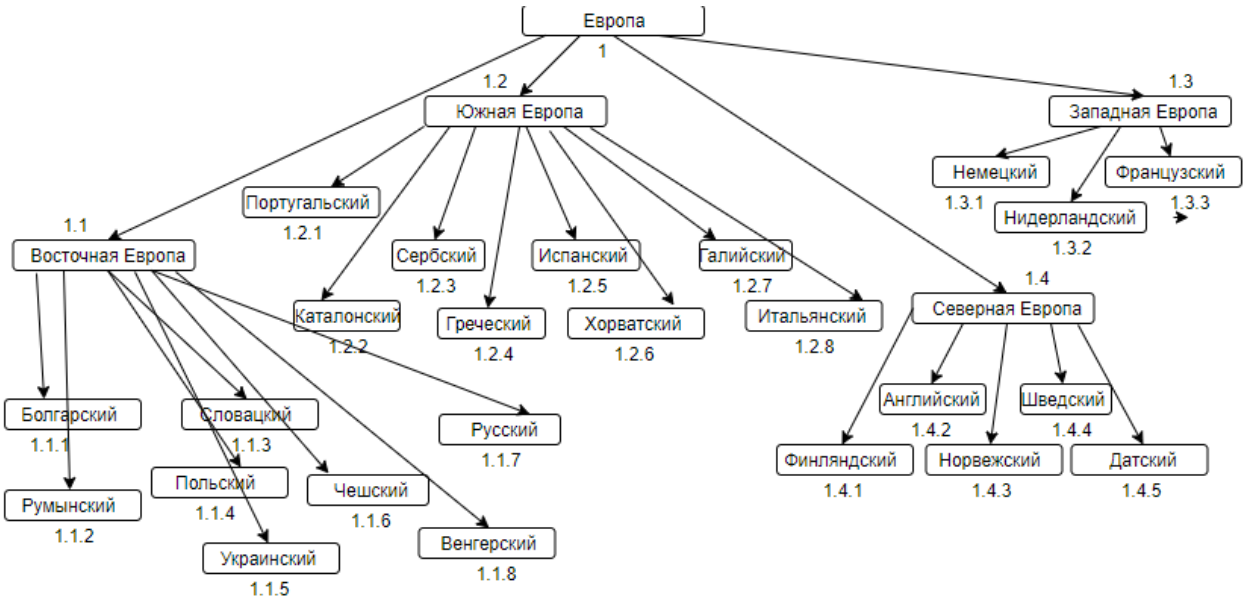

Рисунок 2 – Иерархическая классификация языка (продолжение)

Для поля «media_type» также было выбрано уровневое кодирование. Данное поле принимает всего 3 значения, 2 из которых были отнесены к одной категории, это значения «photo» и c(«photo»,«photo»). Словарь для кодирования представлен в Приложении В.

Таблица 4 – Результаты очистки данных

Параметр	**Исходный набор**	**Набор данных после кодирования**	**Набор, описывающий сущность «Пользователь»**	**Набор, описывающий сущность «Публикация»**
Размер набора - байт	8.68 Мб	8.77 Мб	5.19 Мб	3.7 Мб
Число элементов	16597	16597	16597	16597
Размер элемента (средний)	581 байт	586 байт	304 байт	275 байт
Размер элемента (максимальный)	776 байт	782 байт	366 байт	409 байт
Размер элемента (минимальный)	387 байт	391 байт	242 байт	142 байт
Число фактов в элементе (среднее)	55	55	23	31
Число фактов (слов, единиц языка) в элементе (минимальное)	30	30	17	13
Число фактов (слов, единиц языка) в элементе (максимальное)	80	80	30	50

Таблица 5 – Результаты кодирования данных

Имя поля данных	Способ кодирования (справочника)	Число категорий	Число экземпляров в категории		
			Минимальное	**Медианное**	Максимальное
account_lang	Уровневое кодирование	34	1 (bg, fil, gl, uk, zh-CN)	11	14966 (en)
media_type	Уровневое кодирование	2	3 (c(“photo”,”photo”))	1917 (photo)	14677 (NA)

Если сравнивать исходный набор, который был получен после предобработки данных, и набор, полученный после кодирования признаков, то можно увидеть, что размер набора незначительно, но увеличился, это связано с тем, что изначально каждое значение поля «account_lang» было представлено двумя буквами, а после кодирования длина каждого значения увеличилась.