ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

Челябинский физико-математический журнал. 2018. Т. 3, вып. 2. С. 227–236.

УДК 004.855.5 DOI: 10.24411/2500-0101-2018-13209

АНАЛИЗ ТЕКСТОВ ДЛЯ ПРОГНОЗИРОВАНИЯ ОТТОКА КЛИЕНТОВ ИНТЕРНЕТ-ПРОВАЙДЕРА

А. А. Карякина[a], **Д. С. Ботов**[b]

Челябинский государственный университет, Челябинск, Россия
[a] *suein_i@mail.ru,* [b] *dmbotov@gmail.com*

Прогнозируется отток клиентов на основе данных российского интернет-провайдера. Определены основные этапы и подходы к предварительной обработке текстов комментариев операторов. Предложено использовать для сравнения алгоритмы классификации, такие как логистическая регрессия, метод k-ближайших соседей, градиентный бустинг, наивный байесовский алгоритм. В качестве выборки сформирован массив входных данных из 23 признаков 380 тысяч абонентов. Проведены исправление опечаток с помощью расстояния Дамерау — Левенштейна и лемматизация текстовой информации с последующим преобразованием в вектор признаков с помощью метода TF-IDF и добавлением в модель. Определены основные подходы кодирования категориальных признаков. Построены прогнозные модели. Проведено сравнение результатов исследования на разных классификаторах и сделаны выводы.

Ключевые слова: *прогнозирование, отток клиентов, интернет-провайдер, python, обращения клиентов, классификация, анализ текстов, tf-idf.*

Введение

Рынок телекоммуникаций во всём мире сталкивается с потерей доходов из-за жёсткой конкуренции в борьбе за потенциальных клиентов. Поскольку удержание клиентов в течение длительного времени — это серьёзная проблема, компании вынуждены искать способы использования методов интеллектуального анализа данных и статистических инструментов для определения причины заранее и немедленных ответных действий. К счастью, отрасли телекоммуникаций генерируют и поддерживают большой объём данных. Такой размер информации обеспечивает возможность применения методов интеллектуального анализа данных в телекоммуникационной базе данных.

Среди этих данных объём информации, доступной только в неструктурированной текстовой форме, быстро растёт, поэтому методы классификации играют важную роль в машинном исследовании таких корпусов.

В данной статье проводится исследование использования методов интеллектуального анализа как на обычной информации о клиенте (оплаты, долги и т. д.), так и на текстовой, а также проблем, возникающих при её обработке.

1. Формулировка задачи

Провести проверку возможности применения анализа текстов для прогнозирования вероятности прекращения пользования услугами компании клиентом.

Для этого необходимо:
- проанализировать, агрегировать и собрать данные;
- провести предобработку текстов комментариев операторов в обращениях;
- провести предобработку числовых признаков;
- выбрать алгоритмы для построения моделей, сравнить и определить лучший.

Планируемые результаты проекта со стороны интернет-провайдера, предоставляющего данные, — модель, которая с вероятностью более 50 % определяет абонента на основе его характеристик и комментариев операторов в его обращениях, который закроет договор или прекратит пользоваться услугами компании.

2. Этапы классификации текстов

Отнесение документов к определённым классам на основе их содержимого называется текстовой классификацией. Решение задачи классификации текстов можно разделить на два этапа:

- Преобразование документов — приведение последовательности символов к векторному представлению в пространстве признаков, так как большинство алгоритмов машинного обучения работает именно в нём. Методы преобразования текста в вектор специфичны для каждой задачи и могут зависеть от коллекции документов, типа текста (простой, структурированный) и языка документа.
- Построение модели. Качество классификации также зависит и от алгоритма. Для задачи классификации текстов методы машинного обучения не являются специфичными и применяются также и в других областях [1].

Индексацией документа называется получение вектора признаков для него. Её можно представить в виде двух этапов [2]:

- Получение термов (Term extraction) — на этом этапе применяются методы для поиска и выбора наиболее значимых термов в корпусе документов [3].
- Взвешивание термов (Term weighting) — определение значимости терма для выбранного документа [4].

Под термами понимаются слова или словосочетания, несущие информацию о тематике документа. Термы, или ключевые слова представляют собой краткое описание документа. Существует несколько стандартных способов задания функции взвешивания [5]:
- булевский вес — 1, если слово встречается в документе, иначе — 0;
- tf (term frequency, частота терма) — частотность терма, т. е. как часто терм встречается в документе;
- tf-idf — произведение частоты терма в документе и обратной частотности документов;
- maxstr (maximum strength — максимальная сила) — альтернатива idf.

Широкого распространения булевский вес не получил, так как бинарной информации зачастую оказывается недостаточно для качественной классификации [6].

TF-IDF — простой способ оценить значимость термина для документа относительно всех остальных, поэтому в данной статье будет использована эта мера. Она вычисляется следующим образом [7]:

$$tfidf_{i,j} = tf_{i,j} \ln\left(\frac{N}{df_i}\right),$$

где $tf_{i,j}$ — отношение количества вхождений слова к общему числу терминов документа, df_i — число документов из коллекции, в которых встречается слово, N — число документов в коллекции.

3. Способы токенизации текста

Токенизация — разбиение текста на осмысленные элементы (слова, фразы, символы), называемые токенами [8].

Существует два наиболее распространённых метода токенизации:

- беспорядочное представление документа, также называемое «мешок слов»;
- модель n-gram.

Мешок слов (Bag-of-Words) — это упрощённое представление, используемое для обработки естественного языка и IR. В этой модели текст описывают в виде «мешка» его слов, игнорируя грамматику и даже порядок слов, но сохраняя множественность.

Модель n-gram — это модель представления текстов в виде набора последовательностей, состоящих из N слов. Например, биграммы состоят из двух слов, триграммы — из трёх и т. д.

В данном исследовании будет использоваться модель «мешок слов».

4. Обработка категориальных признаков

Значения факторных признаков для алгоритмов бесполезны: чаще всего категории кодируют разными целыми числами, но использовать на таких данных классические методы машинного обучения, ориентированные на вещественные признаки, нельзя. Для категориальных признаков имеет смысл лишь операция сравнения, поэтому их нужно кодировать. Существует несколько способов [9; 10]: one-hot-кодирование, hashing trick, кодирование интерпретируемыми значениями, проекция на окружность.

One-hot-кодирование подразумевает создание для кодируемого факторного признака N новых признаков, где N — число категорий. Каждый i-й новый признак — бинарный характеристический признак i-й категории.

В пример кодирования интерпретируемыми значениями можно привести замену тарифа абонентской платой. Тогда новый признак упорядочит категории по дороговизне.

Hashing trick (хэширование признаков) — для каждой категории определяется свой уникальный код. Основным преимуществом one-hot перед другими способами кодирования является простота реализации и быстрая модификация. Также преимущество такого кодирования состоит в том, что результат двоичный, а не порядковый, и всё находится в ортогональном векторном пространстве.

В данном исследовании выбран способ one-hot-кодирования.

5. Алгоритмы классификации

Практически все методы машинного обучения применимы к задаче классификации текстов. Чаще используются и показывают лучшие результаты следующие алгоритмы: наивный байесовский классификатор, метод k-ближайших соседей, случайный лес, логистическая регрессия [11].

Случайный лес — устойчивый и гибкий метод, однако это не самый лучший выбор при работе с разрежёнными данными большой размерности, типичным примером которых является «мешок слов» [12].

Регрессионное моделирование — метод, который часто используется для поиска связей между вещественными параметрами. Тем не менее его можно применить и для задач классификации, потому что бинарные значения можно рассматривать как частный случай вещественных, и некоторые регрессионные методы, такие как логистическая регрессия, также естественным образом моделируют и дискретные значения [13]. Также в [12; 14] рекомендуется использовать логистическую регрессию при работе с разрежёнными данными большой размерности, такими как «мешок слов». Поэтому данная работа будет основываться на классификации с использованием именно этого алгоритма со значениями гиперпараметров по умолчанию. Но также будет проведено сравнение с другими моделями: наивный байесовский классификатор, метод k-ближайших соседей, градиентный бустинг.

6. Описание корпуса

В качестве корпуса были взяты данные одного российского интернет-провайдера за 4 месяца около 350 тысяч текущих абонентов, у которых есть безлимитный интернет. Отток клиентов был взят за период в два года, вследствие чего количество ушедших абонентов для выборки удалось увеличить до 30 тысяч. Для каждого из них также были извлечены данные за 4 месяца до расторжения договора. Выбор признаков за данный период показал наилучшие результаты. Для каждого клиента были выбраны следующие признаки:

- вектор признаков, полученный из комментариев операторов в обращениях (более подробное описание проблемы и варианты её решения);
- количество оплат;
- средняя сумма оплат;
- средняя разница в днях между оплатами;
- количество смен тарифов за всё время жизни;
- самый большой долг;
- баланс на конец прошлого месяца;
- скидка на конец прошлого месяца;
- хватает ли на оплату следующего месяца (бинарный);
- был ли отрицательный баланс (бинарный);
- сумма абонентской платы с учётом скидки;
- пользуется ли приложением провайдера (бинарный);
- количество устройств, с которых заходил в приложение провайдера;
- количество ошибок в приложении провайдера;
- есть ли кабельное телевидение (бинарный);
- продолжительность жизни с компанией;
- средний размер трафика в день за последнюю неделю;
- средний размер трафика в день за последние 2 недели;
- суммарный размер трафика в день за последний месяц;
- количество пребываний в административном отключении/приостановлении за всё время жизни;
- самое долгое пребывание в административном отключении/приостановлении за всё время жизни;
- находится ли сейчас в административном отключении/приостановлении или нет;
- самый частый способы оплаты;

- последний способ оплаты.

Обращения — это инциденты, которые создаются при любом взаимодействии с клиентом. В них включаются как звонки и СМС, так и вопросы в чате. Для первого признака было проанализировано около тысячи типов инцидентов. Под типом понимается уникальное сочетание в обращении сервиса, причины и признака. Были исключены инциденты, которые создаются автоматически, а также те, которые содержат в комментариях оценку качества от клиента оператору, так как возможность оценивания клиентом услуг появилась в компании недавно и по таким обращениям недостаточно данных для использования в предсказании. Типы инцидентов, по которым было найдено менее 5 обращений, не рассматривались. Также понадобилось вручную разделить инциденты на те, которые были получены от абонентов, и те, которые создались после звонка оператора. Операторы редко сами звонят абонентам, поэтому в данном случае интерес вызывают только обращения от клиентов. В комментариях инцидентов операторы описывают более развёрнуто проблему клиента, пути её решения и результат. Но данная информация записывается не при каждом обращении. Также есть инциденты, которые создаются после общения клиента с оператором в чате. В таком случае комментарии содержат тексты этих переписок.

Только у 30 % абонентов есть обращения. Всего было взято 290 инцидентов с комментариями, в которых среднее количество токенов — 21.

Конечно, тексты самих разговоров были бы более информативны, однако компания не предоставила таких данных и решила провести пока анализ влияния на предсказание оттока клиентов текстов комментариев операторов.

7. Предобработка текста

Перед использованием машинного обучения на текстовых данных первым делом необходимо перевести комментарии каждого клиента в числовой вектор признаков.

Для начала все комментарии были сгруппированы по клиентам в единый текст, приведённый к нижнему регистру. Затем были отсеяны все символы, которые не являются буквами, например, знаки препинания и стоп-слова. Стоп-слова включают в себя самые распространённые части речи, такие как предлоги, союзы и местоимения [15]. Они исключаются для снижения размерности пространства термов. С помощью регулярных выражений и словаря было произведено удаление незначимых для данной области слов, таких как названия городов, месяцев, дней недели, а также фамилии, имена и отчества клиентов.

В комментариях операторов было замечено значительное количество опечаток и жаргонных слов — около 30 %.

Во время проведения лемматизации текста с помощью библиотеки pymorphy2 были исключены из приведения к нормальной форме жаргонные слова и заменены вручную на схожие по смыслу из словаря.

Для исправления опечаток необходимо было для каждой из них найти в словаре наиболее похожее слово и заменить на него. Для этого использовалось расстояние Дамерау — Левенштейна [16] с модификацией, предложенной в [17]. Необходимо изменить «стоимость» операций вставки и удаления на 2, операции обмена на 1, а операции замены одного символа другим вычислять следующим образом: если клавиши, соответствующие сравниваемым символам, расположены рядом на клавиатуре или сравниваемые символы принадлежат одной фонетической группе, то «стоимость» замены — 1, иначе 2. Была найдена библиотека pyxDamerauLevenshtein [18] с открытым исходным кодом и скорректирована под модификацию.

Перед подачей комментариев каждого клиента модели они были преобразованы в векторы признаков с помощью модуля TfidfVectorizer библиотеки sklearn.

8. Предобработка числовых данных

Пропущенные значения были заполнены нулями с помощью DataFrame.fillna(0), так как это не случайные пропуски и их необходимо учитывать.

Для признаков «Самый частый способ оплаты» и «Последний способ оплаты» выделено четыре способа оплат: в офисе провайдера, банковской картой, через терминал, через какую-либо систему. Они были закодированы с использованием модуля OneHotEncoder и каждый из них преобразовался в матрицу. Пример части одной из них представлен в табл. 1.

Таблица 1

Часть матрицы

Оплата в офисе провайдера	Оплата банковской картой	Оплата через терминал	Оплата через систему
...	...	...	...
0	0	0	1
0	0	1	0
0	1	0	0
0	1	0	0
...	...	...	...

9. Построение моделей и оценка результатов

Данные были разделены на две части с помощью функции train_test_split модуля sklearn.model_selection:

- обучающая выборка (60 %),
- тестовая выборка(40 %).

Приведённое соотношение является наилучшим для данного набора данных и было выбрано опытным путём.

Ушедших клиентов в десять раз меньше, чем активных, поэтому была произведена балансировка классов методом smote из модуля imblearn.over_sampling.

Результаты моделей, полученные с признаками из текстовой информации, представлены в табл. 2. Результаты моделей, полученные без признаков из комментариев, представлены в табл. 3. Так как в исследовании прогнозируется отток клиентов, то в данных таблицах приведены метрики только качества предсказания ушедших абонентов.

Таблица 2

Результаты моделей с учётом комментариев операторов

Алгоритм	Точность	Полнота	F1
Логистическая регрессия	0.66	0.70	0.75
Метод k-ближайших соседей	0.55	0.60	0.69
Градиентный бустинг	0.59	0.60	0.62
Наивный байесовский классификатор	0.43	0.71	0.62

Данные показатели были получены на подобранных оптимальных весах классов. Для класса «текущих клиентов» вес равен 0.3, для «ушедших» — 0.1. Попытка изменить стандартный порог классификаторов не привела к улучшению метрик.

Таблица 3

Результаты моделей без учёта комментариев операторов

Алгоритм	Точность	Полнота	F1
Логистическая регрессия	0.72	0.89	0.80
Метод k-ближайших соседей	0.60	0.85	0.70
Градиентный бустинг	0.66	0.86	0.69
Наивный байесовский классификатор	0.55	0.92	0.69

Лучшие результаты показала логистическая регрессия на данных без признаков из комментариев. Следовательно, можно сделать вывод, что для данной компании малая доля клиентов имеет инциденты, поэтому их количества недостаточно для хорошего обучения модели и тексты из комментариев операторов не являются информативными. Также на представленные показатели значительно повлияло увеличение количества ушедших клиентов.

Код для повторения вышеизложенного эксперимента находится в [19]. Использован язык программирования Python, среды разработки PyCharm и IPython notebook и библиотека scikit learn, pandas (для работы с данными), numpy (для работы с массивами).

10. Заключение

В представленном исследовании данные российского интернет-провайдера подвергнуты методам интеллектуального анализа данных с целью предсказания оттока клиентов. Из проведённого эксперимента можно заметить, что комментарии операторов для данной компании неинформативны для предсказания оттока клиентов.

Был проведён сравнительный анализ моделей, построенных на собранных данных, и выявлен лучший алгоритм для данной задачи — логистическая регрессия, показавшая наиболее высокие значения метрик.

Извлечение ушедших абонентов за два года с целью увеличения примеров положительного класса для лучшего обучения модели и подбор оптимальных весов классов привел к повышению точности результатов.

В будущем планируется получить тексты разговоров обращений клиентов и классифицировать их с сохранением связей между словами.

Список литературы

1. **Агеев, М. С.** Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов : дис. ... канд. физ.-мат. наук / М. С. Агеев. — М., 2004. — 136 с.
2. **Попков, М. И.** Автоматическая система классификации текстов для базы знаний предприятия : магистер. дис. / М. И. Попков. — М.: МГУ им. М. В. Ломоносова, 2014. — 56 с.
3. Terminology extraction [Электронный ресурс]. — URL: https://en.wikipedia.org/wiki/Terminology_extraction (дата обращения: 28.12.2017).
4. **El-Khair, I. A.** Term Weighting [Электронный ресурс]. — URL: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_943 (дата обращения: 27.12.2017).
5. Векторная модель [Электронный ресурс]. — URL: https://ru.wikipedia.org/wiki/Векторная_модель (дата обращения: 28.12.2017).
6. **Токарева, Е. И.** Иерархическая классификация текстов : диплом. работа / Е. И. Токарева. — М.: МГУ им. М. В. Ломоносова, 2010. — 46 с.

7. TF-IDF [Электронный ресурс]. — URL: https://ru.wikipedia.org/wiki/TF-IDF (дата обращения: 29.12.2017).
8. Функция токенизации текста на python [Электронный ресурс]. — URL: http://zabaykin.ru/?p=77 (дата обращения: 29.12.2017).
9. **Дьяконов, А.** Python: категориальные признаки [Электронный ресурс] / А. Дьяконов. — URL: https://alexanderdyakonov.wordpress.com/2016/08/03/python-категориальные-признаки/ (дата обращения: 29.12.2017).
10. **Кравченко, А.** Открытый курс машинного обучения. Тема 6. Построение и отбор признаков [Электронный ресурс] / А. Кравченко. — URL: https://habrahabr.ru/company/ods/blog/ 325422/ (дата обращения: 29.12.2017).
11. **Ikonomakis, M.** Text classification using machine learning yechniques / M. Ikonomakis, S. Kotsiantis, V. Tampakas // WSEAS Transcations on computers. — 2005. — Vol. 4, iss. 8. — P. 966–974.
12. Классификация текстов с помощью мешка слов; руководство [Электронный ресурс]. — URL: http://datareview.info/article/klassifikatsiya-tekstov-s-pomoshhyu-meshka-slov-rukovodstvo/ (дата обращения: 29.12.2017).
13. **Нижибицкий, Е. А.** Обзор алгоритмов классификации документов [Электронный ресурс] / Е. А. Нижибицкий. — URL: http://www.machinelearning.ru/wiki/images/e/ef/NizhibitskyKurs.pdf (дата обращения: 29.12.2017).
14. **Фонарёв, А. Ю.** Машинное обучение с категориальными признаками [Электронный ресурс] / А. Ю. Фонарёв. — URL: http://www.machinelearning.ru/ wiki/images/6/62/2014_517_ FonarevAY.pdf (дата обращения: 29.12.2017).
15. Стоп-слова [Электронный ресурс]. — URL: https://klondike-studio.ru/wiki/stop-slova/ (дата обращения: 29.12.2017).
16. Расстояние Дамерау — Левенштейна [Электронный ресурс]. — URL: https://ru.wikipedia.org/wiki/Расстояние_Дамерау_–_Левенштейна (дата обращения: 29.12.2017).
17. **Zobel, J. Dart, P.** Phonetic String Matching: Lessons from Information Retrieval [Электронный ресурс] / J. Zobel, P. Dart. — URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.18.2138&rep=rep1&type=pdf (дата обращения: 29.12.2017).
18. Исходный код библиотеки pyxDamerauLevenshtein [Электронный ресурс]. — URL: https://github.com/gfairchild/pyxDamerauLevenshtein (дата обращения: 29.12.2017).
19. Исходный код [Электронный ресурс]. — URL: https://github.com/KiraTanaka/Prediction-churn-with-analysis-texts (дата обращения: 29.12.2017).

Поступила в редакцию 31.12.2017
После переработки 04.05.2018

Сведения об авторах

Карякина Алина Александровна, студентка Института информационных технологий, Челябинский государственный университет, Челябинск, Россия; e-mail: suein_i@mail.ru.

Ботов Дмитрий Сергеевич, старший преподаватель кафедры информационных технологий и экономической информатики, Челябинский государственный университет, Челябинск, Россия; e-mail: dmbotov@gmail.com.

Chelyabinsk Physical and Mathematical Journal. 2018. Vol. 3, iss. 2. P. 227–236.

DOI: 10.24411/2500-0101-2018-13209

ANALYSIS OF THE TEXTS FOR PREDICTING THE CHURN OF ISP

A. A. Karyakina[a], D. S. Botov[b]

Chelyabinsk State University, Chelyabinsk, Russia
[a] suein_i@mail.ru, [b] dmbotov@gmail.com

The possibility of forecasting the churn of customers based on the data of the Russian ISP are considered. The basic stages and approaches to the preliminary processing of the texts of operators' comments have been determined. It's offered to use classification algorithms such as the logistic regression, *k*-nearest neighbors method, the gradient boosting, the naive Bayesian algorithm. As a sample, an array of input data from 23 features of 380 000 subscribers was formed. Typos are correcting with using the Dahmerau — Levenshtein distance and lemmatizing of the textual information, and then they are converted into a feature vector using the TF-IDF method and are added to the model. The main approaches of categorical features coding are determined. The forecast models are constructed. Comparison of the results of the study with different classifiers is made and conclusions are drawn.

Keywords: *prediction, clients churn, ISP, python, customers calls, classification, analysis of texts, tf-idf.*

References

1. **Ageev M.S.** *Metody avtomaticheskoy klassifikatsii tekstov, osnovannye na mashinnom obuchenii i zhahiyakh ekspertov* [Methods of automatic text classification based on machine learning and expert knowledge. Thesis]. Moscow, 2004. 136 p. (In Russ.).
2. **Popkov M.I.** *Avtomaticheskaya sistema klassifikatsii tekstov dlya bazy znaniy predpriyatiya* [Automatic text classification system for the knowledge base of the enterprise. Master's Thesis]. Moscow, Lomonosov Moscow State University, 2014. 56 p. (In Russ.).
3. *Terminology extraction.* Available at: https://en.wikipedia.org/wiki/Terminology_extraction, accessed 28.12.2017.
4. **El-Khair I.A.** *Term Weighting.* Available at: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_943, accessed 27.12.2017.
5. *Vector model.* Available at: https://ru.wikipedia.org/wiki/Векторная_модель, accessed 28.12.2017. (In Russ.).
6. **Tokareva E.I.** *Iyerarkhicheskaya klassifikatsiya tekstov* [Hierarchical classification of texts. Graduate work]. Moscow, Lomonosov Moscow State University, 2010. 46 p. (In Russ.).
7. *TF-IDF.* Available at: https://ru.wikipedia.org/wiki/TF-IDF, accessed 29.12.2017. (In Russ.).
8. *Funktsiya tokenizatsii teksta na python* [The function of tokenizing the text in python]. Available at: http://zabaykin.ru/?p=77, accessed 29.12.2017. (In Russ.).
9. **Dyakonov A.** *Python: kategorial'nye priznaki* [Python: categorical features]. Available at: https://alexanderdyakonov.wordpress.com/2016/08/03/python-категориальные-признаки/, accessed 29.12.2017. (In Russ.).
10. **Kravchenko A.** *Otkrytyy kurs mashinnogo obucheniya. Tema 6. Postroyeniye i otbor prizhakov* [Open course of machine learning. Topic 6. Construction and selection of features]. Available at: https://habrahabr.ru/company/ods/blog/325422/, accessed 29.12.2017. (In Russ.).

11. **Ikonomakis M., Kotsiantis S., Tampakas V.** Text classification using machine learning yechniques. *WSEAS Transcations on computers*, 2005, vol. 4, iss. 8, pp. 966–974.
12. *Klassifikatsiya tekstov s pomoshch'yu meshka slov. Rukovodstvo* [Classification of texts using a bag of words. Leadership]. Available at: http://datareview.info/article/klassifikatsiya-tekstov-s-pomoshhyu-meshka-slov-rukovodstvo/, accessed 29.12.2017. (In Russ.).
13. **Nizhybitskiy E.A.** *Obzor algoritmov klassifikatsii dokumentov* [Review of classification algorithms for documents]. Available at: http://www.machinelearning.ru/wiki/images/e/ef/NizhibitskyKurs.pdf, accessed 29.12.2017. (In Russ.).
14. **Fonarev A.Yu.** *Mashinnoye obucheniye s kategorial'nymi priznakami* [Machine learning with categorical features]. Available at: http://www.machinelearning.ru/wiki/images/6/62/2014_517_FonarevAY.pdf, accessed 29.12.2017. (In Russ.).
15. *Stop-slova* [Stop words]. Available at: https://klondike-studio.ru/wiki/stop-slova/, accessed 29.12.2017. (In Russ.).
16. *Rasstoyaniye Damerau — Levenshteyna* [The distance of Dahmerau — Lowenstein]. Available at: https://ru.wikipedia.org/wiki/Расстояние_Дамерау_–_Левенштейна, accessed 29.12.2017. (In Russ.).
17. **Zobel J., Dart P.** Phonetic String Matching: Lessons from Information Retrieval. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.18.2138&rep=rep1&type=pdf, accessed 29.12.2017.
18. *Source code of the library pyxDamerauLevenshtein.* Available at: https://github.com/gfairchild/pyxDamerauLevenshtein, accessed 29.12.2017.
19. *Source.* Available at: https://github.com/KiraTanaka/Prediction-churn-with-analysis-texts, accessed 29.12.2017.

Accepted article received 31.12.2017
Corrections received 04.05.2018