САЙФУЛЛИН М. А., СУЛЕЙМАНОВА А. М. РЕАЛИЗАЦИЯ НАИВНОГО БАЙЕСОВСКОГО КЛАССИФИКАТОРА НОВОСТНЫХ ПУБЛИКАЦИЙ В ФИНАНСОВОЙ СФЕРЕ НА ЯЗЫКЕ ПРОГРАММИРОВАНИЯ RUBY

УДК 004.633.2, ВАК 05.13.18, ГРНТИ 28.29.51

Реализация наивного байесовского классификатора новостных публикаций в финансовой сфере на языке программирования Ruby

М. А. Сайфуллин,
А. М. Сулейманова

Уфимский государственный авиационный технический университет, г. Уфа

В статье описан ход разработки, тестирования и оценки качества работы наивного байесовского классификатора для категоризации новостных материалов финансовой сферы на языке программирования Ruby.

Ключевые слова: бухгалтерия, фильтрация, новости, наивный байесовский классификатор, регулярные выражения, финансы, ruby.

Realization of a Naïve Bayesian Classifier for financial news articles in Ruby programming language

M. A. Saifullin,
A. M. Suleimanova

Ufa State Aviation Technical University, Ufa

This article provides a description of the developing, testing and quality evaluating process of the Naïve Bayesian Classifier purposed for financial news categorization in Ruby programming language.

Keywords: accounting, filtering, news, naïve bayesian classifier, regular expressions, finances, ruby.

Введение

Информированность о готовящихся изменениях в законодательстве, о событиях, происходящих в финансовой сфере, а также о намечающихся в ней трендах – это залог стабильного и успешного функционирования любого финансового подразделения, поскольку оно должно иметь возможность заблаговременно перестраивать рабочие процессы и тем самым избегать неблагоприятных ситуаций с фискальными, правоохранительными органами и контрагентами, которые могут привести к репутационным или материальным потерям.

Однако, контент как информационных ресурсов, так и профессиональных изданий имеет различную тематику, что делает затруднительным для сотрудников финансовых отделов вычленение необходимой им информации, которая бы освещала потенциальные и планируемые реальные изменения, и которую можно было бы использовать в практической профессиональной деятельности.

В решении подобной проблемы может помочь разработка информационной системы, способной собирать и отсеивать релевантные публикации в автоматическом режиме. В основе механизма категоризации новостных материалов такой системы может быть один из самых распространенных и эффективных методов – наивный байесовский классификатор. На рисунке 1 схематически изображен порядок работы классификатора:

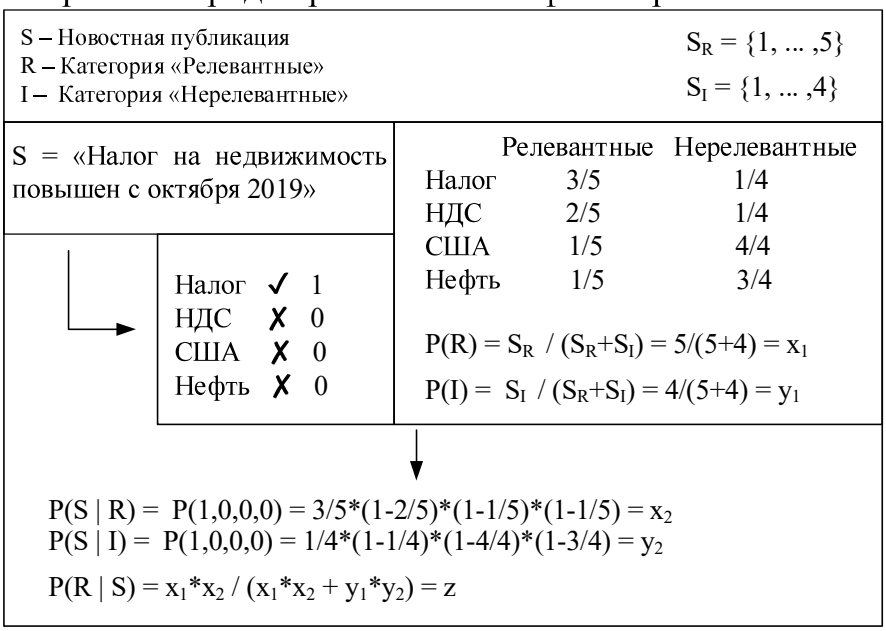

Рисунок 3. Схематический порядок работы классификатора

Полученная величина z – итоговое значение, которое указывает с какой вероятностью публикация S относится к категории R («релевантные»). Аналогичен расчет и для категории I («нерелевантные»).

Оценка вероятности события по данной совокупности признаков, согласно Байесу, основана на произведении условных вероятностей этого события относительно рассматриваемых признаков [1].

Поэтапный процесс разработки классификатора на языке Ruby представлен ниже. Удаленный репозиторий с текстовыми файлами и программным кодом классификатора доступен по адресу: https://github.com/saymirat21/bayes3

Теоретический анализ

1. Сбор исходных данных для анализа

В качестве исходных материалов для разработки классификатора были использованы материалы следующих новостных порталов:

- «Банки.ру» (https://www.banki.ru/);
- «Клерк.ру» (https://www.klerk.ru/);
- «Finanz» (https://www.finanz.ru/);
- «Финмаркет» (http://www.finmarket.ru/);
- «Газета.Ru» (https://www.gazeta.ru/);
- «Главбух» (https://www.glavbukh.ru/);
- «Известия» (https://iz.ru/);

- «Коммерсантъ» (https://www.kommersant.ru/);
- «Журнал – СКБ Контур» (https://kontur.ru/articles);
- «ПРАЙМ» (https://1prime.ru/);
- «РБК» (https://ufa.rbc.ru/);
- «Российская газета» (https://rg.ru/);
- «ВЕДОМОСТИ» (https://www.vedomosti.ru/);
- «Вести Экономика» (https://www.vestifinance.ru/).

В ходе ознакомления с публикациями указанных веб-сайтов проводилась их категоризация на две группы:

- те, которые содержали информацию о потенциальных или реальных изменениях в законодательстве финансовой сферы – в частности, касающихся порядка проведения бухгалтерского учета на предприятиях в общем и учета на предприятиях строительной отрасли;
- прочие новостные материалы, затрагивающие те или иные аспекты сферы экономики в целом.

Одна из основных задач, для решения которой применяют наивный байесовский классификатор – фильтрация спам-сообщений в электронной почте. Как правило, в ходе проектирования классификатора для отсеивания нежелательного контента, сообщения делят на две группы: spam (спам-сообщения – рекламные и прочие подозрительные рассылки) и ham (проверенные, настоящие сообщения).

Подобное правило будет так же сохранено и в данной статье: в дальнейшем, категория релевантных материалов будет обозначаться как «ham», нерелевантные публикации будут относиться к категории «spam».

Фрагмент примерной публикации, относящейся к категории «spam»:

«ЦБ: структурный профицит ликвидности в банковском секторе сократился в сентябре на 0,5 трлн рублей

«В сентябре величина структурного профицита ликвидности снизилась на 0,5 трлн рублей, до 2,9 трлн рублей. Временному снижению профицита в конце месяца способствовало увеличение банками по сравнению с аналогичным периодом предыдущего месяца остатков средств на корсчетах в Банке России», — отмечается в докладе.

<…>

Центробанк повысил прогноз структурного профицита ликвидности на конец 2019 года с 3,4—3,7 до 3,6—3,9 трлн рублей. Пересмотр прогноза связан с размещением бюджетных средств на счета отдельных кредитных организаций.»

Как видно из приведенного выше примера, новость описывает текущие макроэкономические показатели и дает прогноз их развития в дальнейшем: данная информация имеет малую ценность для работников бухгалтерского отдела, поскольку не связана напрямую с областью их трудовой деятельности.

Далее рассмотрим фрагмент релевантной статьи (категория «ham»):

«Фирмы с «обособками» будут по-другому платить НДФЛ

Организации, имеющие несколько обособленных подразделений на территории одного муниципального образования, будут вправе перечислять

НДФЛ в бюджет по месту нахождения одного из таких обособленных подразделений, либо по месту нахождения организации.

<…>

Поправки в п.7 ст. 226 НК и п.2 ст. 230 НК внесены Федеральным законом № 325-ФЗ от 29.09.2019. При этом налоговый агент обязан уведомить о выборе налогового органа не позднее 1-го числа налогового периода налоговые органы, в которых он состоит на учете по месту нахождения каждой «обособки».

Уведомление о выборе налогового органа не подлежит изменению в течение года.»

В результате изучения публикаций новостных порталов была отобрана 1181 статья, 588 из которых были отмечены как релевантные, 593 – отнесены к категории «spam».

2. Первоначальная обработка исходных данных (агрегирование, нормализация и частотный анализ текстов)

Все собранные материалы посредством программного скрипта, находящегося в файле aggregate.rb, были объединены в три текстовых файла: ham_aggregated.txt (все релевантные статьи), spam_aggregated.txt (все нерелевантные статьи) и in_total.txt (статьи обеих категорий).

Перед проведением анализа содержимое статей было нормализовано, что включало в себя:

- замену букв «ё» на «е»;
- снижение регистра букв (замена заглавных букв на строчные);
- удаление латинских символов, цифр, знаков препинания;
- разбиение предложений на массивы слов.

Программный код, представленный в файле words_frequencies.rb (в папке bayes3) создает два файла, в каждом из которых указывается подсчитанное количество появления отдельных слов для каждой из категорий.

В натуральном языке вероятность появления слова сильно зависит от контекста. Алгоритма Байеса использует подход «bag of words model». В этой модели документ представляется в виде мультимножества его слов, игнорируя грамматику и даже порядок слов, но сохраняя множественность, т.е. документ представляется набором отдельных слов, у которых вероятность их появления не зависит друг от друга [2].

На основе изучения наиболее часто встречающихся слов, за исключением служебных частей речи и общеупотребительных слов, по которым не представляется возможным определить тематику публикаций – был сформирован так называемый «bag of words (мешок слов)».

Фрагмент «мешка слов» представлен в таблице 1:

Таблица 1. Часто встречающиеся слова в категориях (фрагмент)

Категория	Слово (однокоренные слова)	Частота появления (соответственно)
spam	Курс (курсов, курса)	115 (48, 31)

	ЦБ (Центробанка, Центробанк, Центробанков, Центробанку)	90 (18, 16, 8, 4)
	Нефть (нефти, нефтяные, нефтяных, нефтью, нефтяной	76 (76, 33, 29, 23, 23)
	Индекс (индексы, индекса, индексов)	51 (24, 11, 9)
	США	174
	Китай (китайские, китайской, китайских, Китае, китайская)	50 (29, 27, 19, 14, 12)
ham	Налог (налога, налоговой, налогов, налогового, налоговый, налогообложения, налоговые)	171 (215, 142, 132, 118, 149, 100, 87)
	Изменения (изменениях, изменениям, изменениями)	151 (14, 7, 3)
	Минфин (Минфина, Минфине, Минфином, Минфину)	142 (101, 35, 20, 8)
	Правительство (правительства, правительством, правительстве	108 (97, 36, 36)
	НДФЛ	137
	Закон (законопроект, закона, законопроекта, закону, законодательства, ФЗ)	143 (114, 89, 51, 46, 41, 77)
	Кодекс (кодекса, кодексе, кодексу, кодексом)	73 (68, 19, 16, 14)
	Ставка (ставку, ставке, ставкам)	81 (53, 39, 8)
	ФНС	101
	Инициатива (инициативу, инициативы, инициативой)	33 (28, 25, 14)
	Поправка (поправки, поправками, поправкам, поправку)	9 (83, 15, 11, 7)

Полная сгенерированная статистика доступна в репозитории в файлах ham_words_frequencies.yaml и spam_words_frequencies.yaml в папке «yaml data».

Как видно из представленной выше таблицы, среди релевантных статей наиболее распространены темы, касающиеся налогообложения и законодательных инициатив, в публикациях категории «spam» затрагиваются темы внешней политики, курсов валют, новостей фондового рынка.

3. Составление таблицы вероятностей классификатора

Исходя из данных таблицы 1 можно сделать вывод о том, что использование инфинитивных форм глаголов и начальных форм существительных недостаточно (из-за наличия множества производных однокоренных слов и альтернативных вариантов написания аббревиатур различных терминов), чтобы наиболее точно оценить частоту появления слов и составить корректную таблицу вероятностей. Поэтому, в дальнейшем, для достижения данных целей в ходе работы будут использоваться regular expressions (регулярные выражения).

В ходе изучения полученной статистики появления в публикациях отдельных слов, был составлен список ключевых слов, представленных в форме регулярных выражений. Часть списка представлена в таблице 2:

Таблица 2. Список ключевых слов (фрагмент)

Ключевое слово (в форме регулярного выражения)
\bцб|центральн.{0,4}\ банк.{0,4}|центробанк.{0,4}|банк\ росс.{0,4}\b
\bминфин.{0,3}|министерств.{0,4}\ финансов\b
\bфсс|фонд социальн.{0,4}\ страхования\b
\bминстро.{0,4}|министерств.{0,4}\ строительства\b
\bналог.{0,12}\b
\bсбор.{0,4}\b
\bзаконопроект.{0,4}\b
\bфз|федеральн.{0,4}\ закон.{0,4}\b
\bндс|налог.{0,4}\ на\ добавленн.{0,4}\b
\bфизлиц.*|физ*.\ лиц.*\b
\bнов.{0,4}|обнов.{0,4}\b
\bиндекс.{0,4}\b
\bсша|америк.*\b
\bсанкц.{0,4}\b
\bувелич.{0,6}\b
\bзапрет.{0,6}\b
\bужесто.{0,6}\b
\bпредлаг.{0,4}|предлож.{0,6}\b

Программный код, находящийся в файле probabilities.rb (папка «bayes3») осуществляет оценку каждого из ключевого слова – подсчитывает сумму его появлений в публикациях и делит ее на общее количество статей в категории. К примеру, если регулярное выражение «\bналог.{0,12}\b» встречается в 15 из 100 статей категории «ham», его вероятность будет равна 15/100=0,15. Расчет для категории «spam» аналогичен. Результатом вычислений является выходной файл – keywords_probs.rb, (находящийся в папке «yaml data»). Фрагмент из полученного расчета в качестве примера представлен в таблице 3:

Таблица 3. Вероятности ключевых слов

Ключевое слово (в форме регулярного выражения)	Вероятность появления в категории	
	ham	spam

\bминфин.{0,3}|министерств.{0,4}\ финансов\b	0.32823	0.10624
\bналог.{0,12}\b	0.67687	0.23946
\bнк|налого.{0,4}\ кодекс.{0,4}\	0.2534	0.06745
\виде.{0,4}|инициатив.{0,4}\b	0.37585	0.17032
\bкадастр.{0,4}\b	0.06633	0.0
\bдолев.{0,4}\ строительств.{0,6}\b	0.06122	0.0
\bувелич.{0,6}\b	0.2398	0.1973
\bновостройк.{0,4}\b	0.01531	0.0
\bэскроу.*\b	0.06293	0.0

После изучения сгенерированного списка, из него были исключены те регулярные выражения, которые имели незначительные вероятности (такие выражения как «\bувелич.{0,6}\b» и «\bдолев.{0,4}\ строительств.{0,6}\b»), и регулярные выражения, у которых разница по модулю между вероятностями мала (например, разница у ключевого слова «\bувелич.{0,6}\b» между значениями ее вероятностей составляет 0,0425).

Полученный набор можно условно разделить на 4 группы: 1) Наименования государственных органов («Минфин», «Минстрой» и др.); 2) Глаголы и существительные, используемые для описания изменений («увеличить», «отменить», «ввести», «предложение», «инициатива» и др.); 3) Профессиональные термины («НДФЛ», «эскроу-счет», «налог», «кадастр» и т.д.); 4) Ключевые слова, часто встречающиеся в нерелевантных публикациях («США», «фондовый индекс» и др.).

Одним из слабых мест наивного Байесовского классификатора является «zero frequency problem» — при наличии нулевых вероятностей у ключевых слов, есть возможность обнуления значений знаменателя, что приведет к делению на ноль — поэтому, вероятности выражений со значением 0.0 были заменены на 0.0001.

Таким образом, в результате был получен набор ключевых слов (регулярных выражений) с наиболее высокими значениями и выраженными разницами в величинах их вероятностей, который находится в файле keywords_probs.yaml в папке «yaml data».

Экспериментальная часть

С помощью программного кода в файле recognize.rb в папке протестируем работу классификатора. В первом случае в качестве релевантной публикации выберем следующую статью (https://kontur.ru/articles/5652):

«Электронная трудовая книжка и новая отчетность в ПФР

1 января 2020 года в России стартует переход на систему электронных трудовых книжек. Это облегчает работу кадровых служб, но и влечет за собой новые обязанности для работодателей. Одна из них — подготовка и сдача в ПФР дополнительного отчета о трудовой деятельности работников — СЗВ-ТД.

<…>

Сейчас в первом чтении Госдума РФ приняла законопроект, в котором сказано, что с 1 января 2021 года работодатели больше не смогут оформлять привычные трудовые книжки для новых сотрудников. В бумажном варианте они сохранятся только у тех работников, кто подаст соответствующее заявление до конца 2020 года. Остальные получат документ на руки. Для всех работников без исключения кадровая служба должна будет вести сведения о трудовой деятельности в электронном виде и подавать их в Пенсионный фонд в виде отчета СЗВ-ТД.

В печатной форме передать отчет СЗВ-ТД смогут только организации с численностью до 25 человек, остальные должны сдавать его в xml-формате.»

Как пример публикации категории «spam» выберем следующую статью (https://kontur.ru/articles/2945):

«Электронный запрос котировок: правила и этапы проведения

Процедура запроса котировок считается одной из самых простых у заказчиков. И не зря — ведь ее можно быстро провести, так как единственным критерием для определения победителя является цена. При этом не нужно требовать финансовое обеспечение заявки и исполнения контракта.

После того, как извещение о проведении электронного запроса котировок размещено и по необходимости были внесены изменения, участники могут подавать свои заявки.

<…>

Заявка на участие в электронном запросе котировок состоит из предложения о товарах, работах, услугах (ТРУ) и отдельного ценового предложения. Эти документы направляются не заказчику, а оператору электронной площадки.»

В результате тестового использования классификатора получаем следующие результаты:

```
saymirat@saymiratpc:~/Desktop/bayes3/bayes3$ ruby recognize.rb
SPAM PROBABILITY: 0.22418201167715976
HAM PROBABILITY: 0.7758179883228403
saymirat@saymiratpc:~/Desktop/bayes3/bayes3$ ruby recognize.rb
SPAM PROBABILITY: 0.9940648044033903
HAM PROBABILITY: 0.005935195596609801
```

Рисунок 4. Результаты тестового применения классификатора

В первом случае вероятность того, что публикация является релевантной, равна примерно 78%, нерелевантной – 22%. Во втором случае публикация классифицирована как «spam» с оценкой в 99%, а вероятность того, что статья относится к категории «ham» – менее процента.

Посредством программного сценария (файл test_all.rb в папке bayes3) оценим качество классификации: применим определение типа публикации ко всем статьям обеих категорий. Результат работы представлен на рисунке 3:

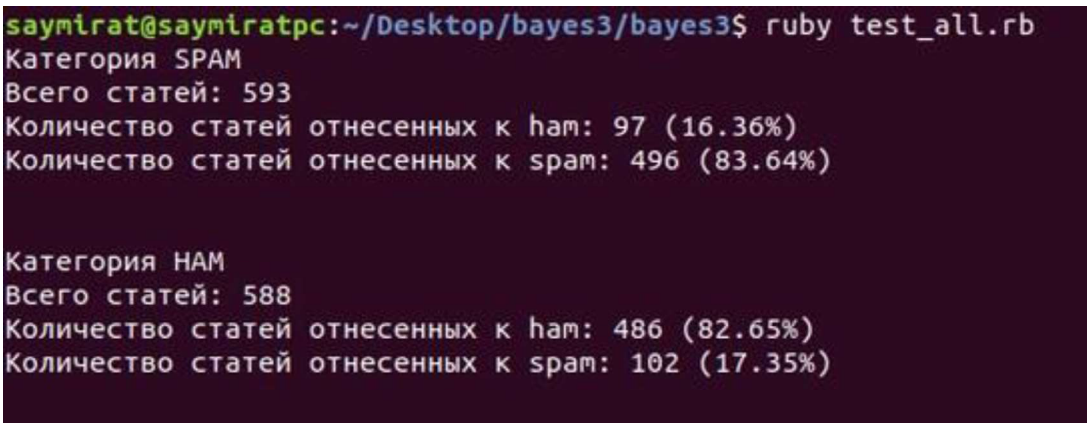

Рисунок 5. Результаты проверки качества работы классификатора

В категории «spam» таковыми были признаны 496 статей, что составляет ≈84% от общего числа публикаций в данной категории. Из 588 релевантных публикаций классификатор распознал только 486, то есть ≈83% из всех статей.

Таким образом, можно сделать вывод о том, процент ошибки в работе разработанного классификатора составляет 16-17%. Данный процент точности является средним для данного типа классификатора – в большей степени точность классификатора для решения данного типа задач зависит от качества сформированного набора ключевых слов («bag of words») и установленных вероятностей их появления в текстах, относящихся к той или иной категории.

Выводы

В ходе разработки наивного классификатора Байеса на языке Ruby был проведен анализ 1181 новостной публикации из 14 информационных источников: материалы были агрегированы, нормализованы, был осуществлен их частотный анализ, на основе которого выделены наиболее часто встречающиеся слова, которые характерны для каждой из категорий.

Составленный перечень слов был переведен в форму регулярных выражений, для которых были рассчитаны вероятностные значения (возможность встретить то или иное слово в статье каждой из двух категории). Полученные вероятности послужили критерием отбора наиболее значимых регулярных выражений, которые впоследствии были использованы для формирования «bag of words (мешка слов)».

После составления необходимых таблиц значений, была проведена тестовая классификация новостных публикаций: выбранным двум статьям разработанный классификатор классы публикаций определил верно; при тестовой оценке отобранных 1181 статей классификатор показал уровень точности равный 81-84%, что является удовлетворительным уровнем ошибки для подобного типа классификаторов.

Список использованных источников и литературы

1. Г. И. Турканов, Е. В. Щепин. Классификатор Байеса для переменного количества признаков // Труды МФТИ. 2016, № 4. С. 8.

2. С.В. Шанов, П.Г. Чупин, А.Ю. Афонин. Применение байесовского классификатора для определения тематики текста // Моделирование, оптимизация и информационные технологии. 2018, №1. С. 133.
3. Классификация текста и наивный Байес. Введение в извлечение информации [Электронный ресурс]: Группа Обработки Естественных Языков Стэнфорда – Url: https://nlp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html (дата обращения: 02.10.2019)
4. Ведомости [Электронный ресурс]: Ведомости – новости бизнеса и финансов, аналитика, смарт-версия газеты – Url: https://www.vedomosti.ru (дата обращения: 02.10.2019)
5. Коммерсант [Электронный ресурс]: Издательский дом «Коммерсантъ» – Url: https://www.kommersant.ru (дата обращения: 28.09.2019)
6. Известия [Электронный ресурс]: Известия – новости политики, экономики, спорта, культуры | IZ.RU – Url: https://www.iz.ru (дата обращения: 18.09.2019)
7. Центральный банк Российской Федерации [Электронный ресурс]: Центральный банк Российской Федерации – Url: https://cbr.ru (дата обращения: 28.09.2019)
8. Минфин России :: Пресс-центр [Электронный ресурс]: Минфин России – Url: https://www.minfin.ru/ru/press-center/ (дата обращения: 28.09.2019)
9. Как работает наивный байесовский алгоритм работает? [Электронный ресурс]: Машинное обучение плюс – Упрощенные руководства на R и Python - Url: https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/ (дата обращения: 28.09.2019)
10. 6 легких шагов к изучению наивного байесовского алгоритма с кодами на Python и R. [Электронный ресурс]: Аналитика Vidhya – Url: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (дата обращения: 28.09.2019)

List of references

1. Turkanov G.I., Scepin E.V. Bayes classifier for a variable number of features. *Trudy MFTI* – Proceedings of Moscow Institute of Physics and Technology, 2016, vol. 8, no. 4, p.8 (in Russian).
2. Shanov S.V., Chupin P.G., Afonin A.Y. Application of the bayesov classifier for the definition of the thematics of the text. *Modelirovanie optimizatsiia i informatsionnye tekhnologii* – Modeling, Optimization and Information Technology, 2018, vol. 6, no. 1, p. 133 (in Russian).
3. Text classification and Naive Bayes. Introduction to Information Retrieval. The Stanford Natural Language Processing Group. Available at: https://nlp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html (accessed: 02.10.2019)
4. Vedomosti – novosti biznesa i finansov, analitika, smart-versia gazety [Vedomosti – business and finance news, analytics, smart newspaper version]. Vedomosti. Available at: https://www.vedomosti.ru (accessed: 02.10.2019)

5. Kommersantie [Bussinesman]: Izdatelskii dom «Kommersantie». Available at: https://www.kommersant.ru (accessed 28.09.2019)
6. Izvestia [News]: Izvestia – novosti politiki, ekonimiki, sporta, kultury | IZ.RU. Available at: https://www.iz.ru (accessed 18.09.2019)
7. TSentralnyi bank Rossiiskoi Federatsii [Central Bank of the Russian Federation]: TSentralnyi bank Rossiiskoi Federatsii. Available at: https://cbr.ru (accessed 28.09.2019)
8. Minfin Rossii :: Press-tsentr [Ministry of Finance :: Press Center]: Minfin Rossii – Available at: https://www.minfin.ru/ru/press-center/ (accessed 28.09.2019)
9. How Naive Bayes Algorithm Works? Machine Learning Plus – Simplified Tutorials in R and Python. Available at: https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/ (accessed: 28.09.2019)
10. 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (accessed at 28.09.2019)