Оценка информативности признаков на основе характеристики тематической значимости при классификации потока новостных сообщений*

В. В. Жебель[I], С-Н. А. Жарикова [II], И. В. Соченков[I]

[I] Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия
[II] Российский университет дружбы народов, Москва, Россия

Аннотация. Статья посвящена оценке качества нескольких методов тематической классификации новостных сообщений. Реализовано несколько известных алгоритмов тематической рубрикации с использованием в качестве признаков различных численных оценок информационной значимости. Рассмотрены классический и предложенный авторами метод определения весов признаков на примере набора данных «20 новостных групп». Представлены полученные результаты экспериментальной апробации системы тематической классификации новостных сообщений, задача которой классифицировать данные на заданные тематические группы. Применение предложенного метода позволяет существенно повысить качество классификации даже с применением базовых методов (мультиномиального наивного байесовского классификатора) до уровня лучших методов в этой области (метод опорных векторов) на эталонном наборе данных.

Ключевые слова: тематический анализ текстов, машинное обучение, характеристика тематической значимости, «20 новостных групп».

DOI 10.14357/20718594190306

Введение

Информационные потоки в современном Интернете весьма интенсивны: ежечасно генерируется и тиражируется значительное число информационных сообщений, затрагивающих различную проблематику. Общее количество информации растет экспоненциально [1]. Одной из ключевых задач является фильтрация и отбор релевантного контента для решения задач мониторинга информационного пространства, построения рекомендательных систем и развития инструментов поддержки принятия решений.

Классификация (т.е. рубрицирование по заранее заданному каталогу) является одним из способов первичной фильтрации и отбора информации в современных информационно-аналитических системах. Алгоритмические и математические основы методов классификации текстов в значительной степени исследованы. Однако показатели качества их работы в значительной степени зависят от специфики наборов данных, на которых решается задача классификации: от жанра текстов (новости, научные публикации, научно-популярные обзоры, патентные документы, диссертации и т.п.) и их объема, а также принятой системы рубрикации.

При этом одним из важных этапов построения классификаторов при использовании классических алгоритмов (в противоположность

* Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 15-29-06082 и № 18-29-16172.

✉ *Соченков Илья Владимирович. E-mail: sochenkov@isa.ru*

подходам на основе глубокого обучения) является выбор информационных признаков. Существует большое количество методов оценки информационной значимости это и классический TF-IDF[2], и статистический χ^2[3], и расстояние Кульбака-Лейблера (Information Gain) [3]. Кроме того, существуют их комбинации, например tf*χ^2, tf*info gain [4].

Данная работа ставит задачу исследования метода оценки информативности признаков на основе характеристики тематической значимости и его применения для повышения качества тематической классификации. В качестве эталонного набора данных для экспериментальной оценки результатов работы алгоритмов классификации новостных сообщений выбран один из стандартных наборов для этой задачи – «20 новостных групп». Проводится исследование классификаторов, построенных с применением классического частотного (TF-IDF) и модифицированного авторами способа оценки информационной значимости лексики, составляющей признаковое пространство для представления текстовых документов.

Выводы исследования могут быть использованы при выборе методов классификации для построения системы первичной фильтрации информации в задачах мониторинга информационного пространства в заданных областях.

1. Оценка информативности признаков в задаче тематической рубрикации

При обработке текстовой информации в современных прикладных системах исследователи ограничиваются рассмотрением отдельных слов в качестве признаков. Для учета разных словоформ в качестве одного признака применяют стемминг (усечение до псевдооснов) или морфологический анализ (приведение к словарной форме). В некоторых случаях для текстов используют более сложные представления в виде графов [5]. Однако в настоящем исследовании мы ограничимся сопоставлением нескольких способов оценки информативности отдельных слов. Как будет показано далее, эти оценки могут применяться и для более сложных конструкций (например, устойчивых словосочетаний или фраз).

Рассмотрим известный способ оценки информационной значимости слов на основе TF-IDF[2]. Данный метод широко применяется для определения значимости признаков и расчета расстояний «документ - категория» и «документ-документ».

Значение частоты встречаемости слова w_i в тексте d

$$tf(w_i, d) = \frac{Wc(w_i, d)}{Awc(d)}, \quad (1)$$

где $Wc(w_i, d)$ - количество употреблений слова w_i в тексте d;

$Awc(d)$ - количество употреблений всех слов в тексте d.

Часто употребляемые в тексте слова (исключая служебные) характеризуют его тематическую направленность. Чтобы отделить служебную лексику вводят величину IDF – инверсную частоту встречаемости слова в коллекции:

$$idf(w_i) = \log_2 \frac{N}{N_{wc}(w_i)}, \quad (2)$$

где N - количество текстов в коллекции;

$N_{wc}(w_i)$ – количество текстов, в которых слово w_i встречается хотя бы один раз.

Инверсная частота встречаемости слова позволяет определять часто встречающиеся слова (почти в каждом документе один или более раз), которые малоинформативны в рассматриваемой коллекции («стоп-слова» в терминологии поисковых систем).

Классическая оценка значимости слова в заданном документе d коллекции является произведением частоты встречаемости слова в документе на его инверсную частоту в коллекции:

$$tf\_idf(w_i, d) = tf(w_i, d) \times idf(w_i). \quad (3)$$

Данный метод взвешивания слов широко используется в качестве базового, однако он подвержен ряду недостатков.

1. Величина TF-IDF чувствительна к изменению длин текстов (в коротких текстах множитель TF значительно больше для всех слов, нежели в длинных текстах): TF-веса слов превалируют над сомножителем IDF в коротких текстах, и обратно – сомножитель IDF в значительной степени определяет общий вес в длинных текстах. Негативное влияние этого эффекта в задачах информационного поиска и классификации текстов привело к созданию различных модификаций – BM25[6], BM25F[7], BM25+ [8].

2. Величина TF-IDF не учитывает дополнительную информацию о частотах слов в документах, соотнесенных с рубриками некоторого тематического классификатора, т.е. не позволяет учесть дополнительную информацию о тематической принадлежности документов.

Для устранения первого недостатка рассмотрим величину LTF («логарифмическая TF»):

$$ltf(w_i, d) = \log_{Awc(d)+1}(Wc(w_i, d) + 1). \quad (4)$$

Эта величина положительно определенная и не превосходит 1.

Для нормализации IDF рассмотрим следующее определение:

$$nidf(w_i) = \frac{\log_2 \frac{N}{N_{wc}(w_i)}}{\log_2 N}. \quad (5)$$

Если $\mathrm{N_{wc}(w_i)} = 0$, то по определению положим $\mathrm{nidf(w_i)} = 1$.

С учетом определенных величин нормированный вариант TF-IDF задается формулой:

$$ltf\_nidf(w_i, d) = ltf(w_i, d) \times nidf(w_i). \quad (6)$$

Для учета тематической информации рассмотрим характеристику тематической значимости (ХТЗ)[9]. Величина, определяемая формулой, показывает изменение информативности слова $\mathrm{w_i}$, при условии, что нам известно, в каком количестве документов из A, это слово встретилось хотя бы один раз:

$$\Delta idf(w_i, A, C) = \log_2 \frac{|C \backslash A|}{Cnt(w_i, C \backslash A)} - \log_2 \frac{|A|}{Cnt(w_i, A)}, \quad (7)$$

где A – множество документов некоторой выбранной рубрики;

C – множество документов всей коллекции (все документы из всех рубрик);

$|.|$ – мощность множества;

$Cnt(w_i, K)$ - количество документов из множества K, в которых слово w_i встречается хотя бы один раз.

Величина Δidf тем больше, чем чаще слово встречается в определенной рубрике, но при этом реже в остальных рубриках. Аналогичный принцип был исследован в работе [10].

Величину, заданную формулой (7), мы будем рассматривать в нормированном на 1 варианте:

$$\Delta idf^*(w_i, A, C) = \frac{1 - \log_2 \frac{|A|}{Cnt(w_i, A)}}{\log_2 \frac{|C \backslash A|}{Cnt(w_i, C \backslash A)}}. \quad (8)$$

Заметим, что эта величина может быть отрицательной, так как вычитаемое в формуле (7) может оказаться больше уменьшаемого (в отличие от подхода, предложенного в [9]). При этом отрицательные значения $\Delta idf^*(w_i, A, C)$ свидетельствуют о том, что слово w_i не является информативным признаком при отнесении документа с этим словом к рубрике документов A, но при этом является значимым, если относить этот документ к множеству $C \backslash A$.

С учетом вышесказанного, оценку информационной значимости слова – характеристику тематической значимости (Topical Importance Characteristic) определим формулой:

$$TIC(w_i, d, A, C) = ltf(w_i, d)\Delta idf^*(w_i, A, C). \quad (9)$$

В качестве замечания укажем, что в вышеприведенных определениях величина $Cnt(w_i, K)$, где K – некоторое подмножество документов, может быть заменена на $Cnt(w_i, K, th)$ с тем же смыслом, но th – суть минимальный пороговый вес (например, величины $ltf(w_i, d)$). То есть $Cnt(w_i, K, th)$ – количество документов во множестве K, что слово w_i удовлетворяет неравенству:

$$ltf(w_i, d) \geq th. \quad (10)$$

Это определение позволяет отфильтровать случайно встретившиеся («очень редкие») слова в документе d при расчете инверсных частот встречаемости и равнозначно усилению исходного определения величины $Cnt(w_i, K)$ заменой слов «встречается хотя бы один раз» на «удовлетворяет соотношению 10».

Введенная величина ХТЗ (формула (9)) является средством фильтрации признакового пространства и отсечения малоинформативных слов при работе с векторным представлением текста в виде «мешка слов» (bag-of-words).

Далее мы экспериментально оценим качество решения задачи классификации с использованием этой величины в виде значений признаков и сравним получаемые результаты с другими классификаторами.

2. Анализ качества тематической рубрикации новостных сообщений с применением метода оценки информативности признаков

В качестве эталонного набора данных для проведения экспериментального исследования выбран корпус «20 новостных групп» («20-Newsgroups») [11]. Этот набор данных содержит 18821 новостное сообщение, отнесенное к одной из двадцати рубрик, и является одним из эталонов для оценки качества алгоритмов текстовой классификации. Этот набор данных содержит относительно короткие тексты. Наряду с новостной спецификой, он значительно «зашумлен», что снижает качество результатов работы методов классификации без предварительной «очистки» данных [12]. Поэтому проведение экспериментов на этом наборе данных с применением ХТЗ для оценки значимости признаков представляется оправданным.

В экспериментальной части исследования мы применяем пятикратную кросс-валидацию с разбиением набора данных в соотношении 80% (обучающая выборка) на 20% (контрольная выборка).

В качестве инструмента для обработки данных, реализации классификаторов и оценки качества их работы используется библиотека sklearn [13]. Для первичной обработки данных используются компоненты библиотеки CountVectorizer и TfidfVectorizer. На этом этапе мы дополнительно используем следующую эвристику для повышения разреженности матрицы «слова Х документы», отфильтровывая частотные «стоп-слова» (встречаются в большом количестве документов) и редкие слова (обычно – опечатки или неинформативные служебные элементы, например, адреса электронной почты): $0{,}005|C| < Cnt(w_i, C) < 0{,}22|C|$.

На подготовленных этим способом данных мы провели по три серии экспериментов для двух базовых алгоритмов классификации: мультиномиального наивного байесовского классификатора (MultinomialNB) и классификатора на основе случайного леса (`RandomForest`). В сериях экспериментов сравнивалось качество классификации в зависимости от выбранного метода определения информативности слов: с помощью классического TF-IDF (формула (3)) и модифицированных (формулы (6) и (9)) подходов.

Алгоритм MultinomialNB реализует наивный алгоритм Байеса для полиномиально распределенных данных и является одним из базовых решений для классификации текстов. Метод RandomForest заключается в построении множества деревьев принятия решений, каждое из которых представляет собой набор иерархических условий классификации объектов к тому или иному классу на основе независимых признаков. В эксперименте количество деревьев было эмпирически подобрано равным 100 для достижения наилучшего результата.

Оба классификатора не являются лучшими по качеству в задачах текстовой классификации, особенно на зашумленных данных и в условиях относительно небольших обучающих выборок (к таковым можно отнести и выбранный набор данных «20 новостных групп»). Однако оба алгоритма чувствительны к качеству отобранных информативных признаков, что позволяет проанализировать и сопоставить предложенные методы.

Решаемая задача относится к непересекающейся замкнутой классификации (каждый документ принадлежит одной и только одной рубрике), поэтому в качестве метрики качества может быть выбрана аккуратность классификации (Accuracy). Мы будем использовать макроусреднение этой метрики.

В Табл. 1 приведены результаты сопоставления качества работы классификаторов на контрольной выборке по четырем рубрикам, каждая из которых хорошо обособлена тематически от остальных:

- 'sci.space',
- 'soc.religion.christian',
- 'talk.politics.guns',
- 'comp.windows.x'.

Ожидаемо у классификаторов не возникает трудностей с определением рубрик, так как они неплохо разделяются за счет различного состава лексики. При этом метод, склонный к переобучению при работе с текстовой информацией (RandomForest), демонстрирует качество на уровне метода с большей обобщающей способностью в общем случае (MultinomialNB).

Ситуация усложняется, если рассмотреть тематически близкие рубрики. Табл. 2 содержит результаты сопоставления качества работы классификаторов на контрольной выборке по

четырем рубрикам, каждая из которых принадлежит к компьютерной тематике:

- 'comp.os.ms-windows.misc',
- 'comp.sys.ibm.pc.hardware',
- 'comp.sys.mac.hardware',
- 'comp.windows.x'.

Для этого набора категорий наблюдается снижение аккуратности классификации. На этой задаче оба метода при использовании нетематических признаков (tf_idf и ltf_nidf) демонстрируют худшие результаты по сравнению с взвешиванием на основе ХТЗ (TIC), которая обеспечивает на 10% лучшую аккуратность классификации в методе MultinomialNB. С другой стороны, RandomForest «отбирает» информативные признаки на этапе построения деревьев решений, поэтому здесь влияние ХТЗ не наблюдается, а все три способа взвешивания слов дают одинаковые результаты («способность к обобщению» не увеличивается).

Приведем результаты сопоставления качества работы классификаторов на полном наборе рубрик (Табл. 3). Классификаторы на основе ХТЗ работают с более высоким качеством, нежели с альтернативными способами вычисления информативности признаков. MultinomialNB демонстрирует стабильные результаты на полном наборе рубрик. RandomForest с применением ХТЗ также имеет лучшую обобщающую способность, нежели с использованием нетематических оценок. Нормированный вариант TF-IDF (ltf_nidf)

Табл. 1. Аккуратность классификации по тематически обособленным рубрикам, %

Номер испытания	**MultinomialNB**			**RandomForest**		
	tf_idf	ltf_nidf	TIC	tf_idf	ltf_nidf	TIC
1	97,43	97	99,79	95,29	94,86	97,43
2	97,43	97,86	99,98	96,36	96,57	98,5
3	97,86	98,29	99,79	95,72	94,86	98,07
4	97,43	97,22	99,57	95,93	95,72	96,15
5	97,22	97	98,72	95,29	95,07	99,14
СРЕДНЕЕ	97,47	97,47	**99,57**	95,72	95,41	97,85

Табл. 2. Аккуратность классификации по тематически близким рубрикам, %

Номер испытания	**MultinomialNB**			**RandomForest**		
	tf_idf	ltf_nidf	TIC	tf_idf	ltf_nidf	TIC
1	86,41	86,62	96,82	85,35	86,41	85,35
2	87,69	85,77	95,97	87,05	87,05	84,71
3	86,2	83,44	94,9	84,71	85,56	86,62
4	85,77	84,5	94,48	85,14	84,93	82,8
5	84,5	82,59	95,97	86,2	84,93	88,32
СРЕДНЕЕ	86,11	84,58	**95,63**	85,69	**85,78**	85,56

Табл. 3. Аккуратность классификации по полному множеству рубрик, %

Номер испытания	**MultinomialNB**			**RandomForest**		
	tf_idf	ltf_nidf	TIC	tf_idf	ltf_nidf	TIC
1	82,99	80,73	94,87	81,04	81,93	87,89
2	83,47	81,44	95,23	81,44	82,15	88,56
3	83,56	82,55	97,70	81,75	81,57	89,31
4	80,95	79,14	93,73	80,38	80,73	88,95
5	82,59	81,09	95,23	80,11	80,91	88,11
СРЕДНЕЕ	82,71	80,99	**95,35**	80,94	81,468	**88,56**

на этой задаче не дает выигрыша перед ненормированным TF-IDF (tf_idf). Впрочем, это может быть объяснено тем, что влияние различий длин документов не проявляется на выбранном наборе данных, а специфика текстов такова, что вклад величины TF (в классическом варианте) в общую величину TF-IDF позволяет получить более информативные слова в этих текстах.

3. Связанные исследования и приложения

Полученные результаты соответствуют и сопоставимы с уровнем исследований в области текстовой классификации. Например, в работе [14] авторы используют вычислительно более сложный алгоритм: метод опорных векторов с линейным ядром, применяемый к усредненному векторному представлению лексики текста на основе взвешенных векторов, полученных с помощью инструмента моделирования дистрибутивной семантики word2vec. В этом методе используется взвешивание по модифицированной схеме TF-IDF (сходной с BM-15), и на том же наборе данных получено значение аккуратности, равное 89,7%.

В работе [15] исследователи применяют комбинированный подход к выявлению информативных признаков для тематических категорий. Наиболее близкими к предложенному в настоящей работе методу являются энтропийный подход и количество полезной информации (Information Gain). Авторы достигают аккуратности классификации порядка 80% при использовании усеченного признакового пространства размерностью в 1000 наиболее информативных признаков.

В работе [16] было рассмотрено сразу большое количество методов взвешивания (различные производные tf*idf, BM25, RF, LR, …) для того же набора данных [11], при этом результаты аккуратности оказались значительно ниже продемонстрированных нами.

Задача классификации является не единственной в области обработки текстовой информации, которая требует выявления значимых ключевых элементов текста. Предложенный подход может применяться для выделения составных конструкций, типичных для текстов, относящихся к заданным категориям, жанрам и т.п. К близким задачам относятся тематическая кластеризация коллекций текстов, а также фильтрация потока документов (например, новостного). Примером решения задачи первичной фильтрации информации может служить система сфокусированного сбора информации (краулинга) [17]. Специфика этой задачи заключается в необходимости быстрого анализа загружаемых документов (без возможности выполнения полного лингвистического анализа текста, требующего значительных вычислительных затрат) с приоритетом по полноте над точностью. Предложенный метод оценки информативности может быть применен и в этом случае, причем в качестве признаков могут выступать «токены», составляющие текст документа (без нормализации и стемминга).

Заключение

В статье предложены и экспериментально проверены подходы к определению значимости признаков для документов и рубрик в задаче тематической классификации. Предложенный подход позволяет учесть тематическую принадлежность документов при оценке информационной значимости элементов текстов в задачах тематической классификации. Анализ результатов показывает, что применение предложенного метода позволяет существенно повысить качество классификации даже с применением базовых методов (мультиномиального наивного байесовского классификатора) до уровня лучших методов в этой области (метод опорных векторов) на стандартном наборе данных.

Результаты исследования использованы в прототипе системы для исследования информационного пространства, связанного с арктической зоной.

Литература

1. Huberman B. A., Adamic L. A. Internet: growth dynamics of the world-wide web //Nature. 1999. Т. 401. №. 6749. – P. 131.
2. Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. – Carnegie-mellon univ pittsburgh pa dept of computer science. 1996. № CMU-CS-96-118.
3. Caropreso M.F., Matwin S., Sebastiani F. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In Text Databases and Document Management: Theory and Practice, A. G. Chin, ed. Idea Group Publishing, Hershey, PA. 2001.– P.78–102.

4. Ying Liu, Han Tong Loh, Aixin Sun Imbalanced text classification: A term weighting approach // Expert Systems with Applications.Volume 36, Issue 1. 2009.– P. 690-701.
5. Sonawane S. S., Kulkarni P. A. Graph based representation and analysis of text document: A survey of techniques //International Journal of Computer Applications. 2014. Т. 96. № 19.
6. Robertson S. E. et al. Okapi at trec-3 proceedings of the third text retrieval conference. – TREC. 1994.
7. Robertson S. et al. The probabilistic relevance framework: BM25 and beyond //Foundations and Trends® in Information Retrieval. 2009. Т. 3. № 4. – P. 333-389.
8. Lv Y., Zhai C. X. Lower-bounding term frequency normalization //Proceedings of the 20th ACM international conference on Information and knowledge management. – ACM. 2011. – P. 7-16.
9. Suvorov R., Sochenkov I., Tikhomirov I. Method for pornography filtering in the web based on automatic classification and natural language processing //International Conference on Speech and Computer. – Springer, Cham. 2013. – P. 233-240.
10. Martineau J. et al. Delta TFIDF: An Improved Feature Space for Sentiment Analysis //Icwsm. 2009. Т. 9. – P. 106.
11. The Twenty Newsgroups. [Электронный ресурс]. URL: http://qwone.com/~jason/20Newsgroups/ (дата обращения 10.02.2018 г.).
12. Albishre K., Albathan M., Li Y. Effective 20 newsgroups dataset cleaning //Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on. IEEE. 2015. Т. 3. – P. 98-101.
13. Prettenhofer P. et al. Scikit-learn: Machine Learning in Python. 2016.
14. Lilleberg J., Zhu Y., Zhang Y. Support vector machines and word2vec for text classification with semantic features //Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on. IEEE. 2015. – P. 136-140.
15. Tang B. et al. A Bayesian classification approach using class-specific features for text categorization //IEEE Transactions on Knowledge and Data Engineering. 2016. Т. 28. № 6. – P. 1602-1606.
16. Mendoza, M. A new term-weighting scheme for naïve Bayes text categorization // International Journal of Web Information Systems. 2012. Vol. 8. № 1. – P. 55-72.
17. Devyatkin D., Shelmanov A. Text Processing Framework for Emergency Event Detection in the Arctic Zone //International Conference on Data Analytics and Management in Data Intensive Domains. – Springer. Cham. 2016. – P. 74-88.

Feature Selection for Text Classification of a News Flows based on Topical Importance Characteristic

V. V. Zhebel [I], S.-N. A. Zharikova [II], I. V. Sochenkov [I]

[I] Federal Research Center "Computer Science & Control" of the Russian Academy of Sciences Moscow, Russia
[II] Peples' Friendship University of Russia, Moscow, Russia

Abstract. The paper presents an approach for ranking the most valuable features for text classification task. The introduced Topical Importance Characteristic leverages the feature selection method comprising the information about the distributions of words or phrases among the topics. We compare this method to well-known TF-IDF approach and use the introduced word-ranking scheme in two classifiers: Random Forrest and Multinomial Naïve Bayes. The Accuracy of classification results was tested in the "20-Newsgroups" dataset. The developed approach outperforms TF-IDF-based methods and matches the Accuracy achieved by the more powerful state of the art approaches such as SVC on the same dataset.

Keywords: topical text classification, machine learning, topical importance characteristic, 20-Newsgroups.

DOI 10.14357/20718594190306

References

1. Huberman B. A., Adamic L. A. Internet: growth dynamics of the world-wide web //Nature. – 1999. – Т. 401. – №. 6749. – С. 131.
2. Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. – Carnegie-mellon univ pittsburgh pa dept of computer science, 1996. – №. CMU-CS-96-118.
3. CAROPRESO, M. F., MATWIN, S., AND SEBASTIANI, F. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In Text Databases and Document Management: Theory and Practice, A. G. Chin, ed. Idea Group Publishing, Hershey, PA, P.78–102.
4. Ying Liu, Han Tong Loh, Aixin Sun Imbalanced text classification: A term weighting approach // Expert Systems with Applications.Volume 36, Issue 1. 2009. pp. 690-701

5. Sonawane S. S., Kulkarni P. A. Graph based representation and analysis of text document: A survey of techniques //International Journal of Computer Applications. – 2014. – Т. 96. – № 19.
6. Robertson S. E. et al. Okapi at trec-3 proceedings of the third text retrieval conference. – TREC, 1994.
7. Robertson S. et al. The probabilistic relevance framework: BM25 and beyond //Foundations and Trends® in Information Retrieval. – 2009. – Т. 3. – №. 4. – P. 333-389.
8. Lv Y., Zhai C. X. Lower-bounding term frequency normalization //Proceedings of the 20th ACM international conference on Information and knowledge management. – ACM, 2011. – P. 7-16.
9. Suvorov R., Sochenkov I., Tikhomirov I. Method for pornography filtering in the web based on automatic classification and natural language processing //International Conference on Speech and Computer. – Springer, Cham, 2013. – P. 233-240.
10. Martineau J. et al. Delta TFIDF: An Improved Feature Space for Sentiment Analysis //Icwsm. – 2009. – Т. 9. – С. 106.
11. The Twenty Newsgroups. [Электронный ресурс]. URL: http://qwone.com/~jason/20Newsgroups/ (дата обращения 10.02.2018 г.)
12. Albishre K., Albathan M., Li Y. Effective 20 newsgroups dataset cleaning //Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on. – IEEE, 2015. – Т. 3. – P. 98-101.
13. Prettenhofer P. et al. Scikit-learn: Machine Learning in Python. – 2016.
14. Lilleberg J., Zhu Y., Zhang Y. Support vector machines and word2vec for text classification with semantic features //Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on. – IEEE, 2015. – P. 136-140.
15. Tang B. et al. A Bayesian classification approach using class-specific features for text categorization //IEEE Transactions on Knowledge and Data Engineering. – 2016. – Т. 28. – №. 6. – P. 1602-1606.
16. Mendoza, M. A new term-weighting scheme for naïve Bayes text categorization // International Journal of Web Information Systems, 2012. Vol. 8 No. 1, pp. 55-72.
17. Devyatkin D., Shelmanov A. Text Processing Framework for Emergency Event Detection in the Arctic Zone //International Conference on Data Analytics and Management in Data Intensive Domains. – Springer, Cham, 2016. – P. 74-88.