УДК 81’322.+81’33+811.58
ББК Ш.11+Ш171
DOI: 10.14529/ling160304

ИСПОЛЬЗОВАНИЕ ТЕОРИИ МНОЖЕСТВ В СРАВНИТЕЛЬНО-ЛОГИЧЕСКИХ МЕТОДАХ ВЫДЕЛЕНИЯ ТЕКСТОВ НА ИСЛАМСКУЮ ТЕМАТИКУ В ПРОЦЕССЕ МОНИТОРИНГА СЕТЕВЫХ РЕСУРСОВ

Б.Г. Фаткулин
Южно-Уральский государственный университет, г. Челябинск

Выделение текстов исламского содержания в сетевых ресурсах может проводиться с помощью сравнительно-логических методов «проверка на вхождение» и «пересечение множеств». Для какого-либо языка предварительно собирается корпус прецедентных текстов на исламскую тематику. Из этого корпуса при помощи статистических методов извлекается множество ключевых понятий. Список выделенных ключевых слов представляется в виде множества. Сформированное множество ключевых понятий может быть представлено в виде базы данных и в дальнейшем используется в качестве эталонного множества А. Тексты, представляющие интерес для экспертизы, представляются в виде множества понятий Б. Множество А сравнивается с множеством А на предмет пересечения. Наличие в множестве Б элементов множества А, характер и степень пересечения двух множеств позволяет идентифицировать текст, представляющий интерес для экспертизы. Язык программирования Python предоставляет широкие возможности для работы со строками, кортежами, словарями и множествами. В процессе мониторинга сетевых ресурсов и поиска текстов на исламскую тематику необходимо использовать методы, изложенные в статье, а также базы данных с эталонными списками ключевых слов.

Ключевые слова: информационный экстремизм, судебная лингвистическая экспертиза, сетевые ресурсы, мониторинг сетевых ресурсов, сравнительно-статистические методы, язык Python, множества, лингвистические ресурсы, прецедентные тексты, ислам.

Повышение уровня информатизации современного общества, расширение массива печатной и электронной продукции приводят к количественному росту и значительному содержательному и структурному усложнению спорных и конфликтогенных текстов, некоторые из которых могут носить экстремистский характер.

Экстремизм в своем юридическом значении определен в статье 1 Федерального Закона № 114-ФЗ «О противодействии экстремистской деятельности», в котором дан достаточно широкий круг явлений, подпадающих под данное понятие. По мнению О.С. Жуковой, Р.Б. Иванченко, В.В. Трухачева, «Информационный экстремизм — это деятельность, связанная с: а) созданием, хранением и (или) распространением информации, содержащей предусмотренные законом признаки экстремистской деятельности; б) использованием информации, обрабатываемой компьютером, компьютерной системы и (или) компьютерной сети, осуществляемым в целях воздействия на принятие решения органами государственной власти, органами местного самоуправления или международными организациями, сопряженным с различными формами психического или опосредованного физического насилия (кибертерроризм); в) использованием информации, оказывающей деструктивное воздействие на психику людей, не осознаваемым ими» [1].

Информационный экстремизм неотделим от понятия информационной среды. Информационная среда содержит в себе информационные и сетевые ресурсы — отдельные документы и массивы документов в библиотеках, архивах, фондах, банках данных и других информационных системах. Сетевой ресурс — это часть информации, к которой может быть осуществлен удаленный доступ с другого компьютера, обычно через локальную компьютерную сеть или посредством корпоративного интернета, как если бы ресурс находился на локальной машине. Именно в отношении таких ресурсов имеется опасность их массового распространения по электронным библиотекам, on-line магазинам, сети Интернет.

Сетевые ресурсы представлены в цифровой форме, поэтому рутинные простейшие процедуры мониторинга сетевых ресурсов могут быть осуществлены с использованием методов прикладной лингвистики. Мониторинг сетевых ресурсов, проводимый в целях первичной идентификации, поможет выявить экстремистские материалы на этапе их порождения. Результаты мониторинга будут иметь процессуальные последствия только в случае соответствующей его оценки в рамках судебной экспертизы. Правила назначения и проведения судебной экспертизы установлены Уголовно-процессуальным кодексом РФ, Гражданским процессуальным кодексом РФ и Федеральным законом «О государственной судебно-экспертной деятельности в Российской Федерации» от 31 мая 2001 г. № 73-ФЗ.

Судебная лингвистическая экспертиза (разновидность судебной экспертизы) определяется как «процессуально регламентированное экспертное лингвистическое исследование устного и (или) письменного текста, завершающееся составлением письменного заключения по вопросам, разрешение которых требует применения специальных познаний».

Целью досудебной профилактической экспертизы информации является отсечение информационных ресурсов потенциально опасной направленности. Закономерна последующая судьба информационных материалов, признанных судом экстремистскими. В соответствии со ст. 13 Федерального закона «О противодействии экстремистской деятельности» все экстремистские материалы заносятся в единый федеральный список, который размещается в Интернете на официальном сайте Министерства юстиции РФ.

Судебная лингвистическая экспертиза опирается на доказательства, которыми как правило являются выводы экспертов о том, что тексты сообщений, высказывания в соцсетях, материалы, распространяемые на интернет-страницах содержат элементы экстремизма. Таким образом мониторинг информационной среды является одним из эффективных методов профилактики информационного экстремизма [2].

Идентификация характера текстов может являться одной из целей мониторинга сетевых информационных ресурсов. Мониторинг сетевых ресурсов может осуществляться как вручную, так и с помощью специальных средств. Специально разработанные компьютерные программы помогают преодолеть рутину в ходе обработки цифрового контента, т. е. оцифрованных текстов.

В РФ мониторинг информационных ресурсов в сети Интернет направлен в основном на выявление экстремистских материалов на государственном (русском) языке. Поэтому в таком мониторинге участвуют как правило филологи-русисты. Однако многонациональный состав населения РФ и глобализация привели к тому, что проповедники экстремизма используют для своих целей не только русский язык, но и языки субъектов федерации, а также иностранные языки, в том числе восточные [3].

В ходе мониторинга информационных ресурсов и их идентификация используются различные методы прикладной лингвистики. Настоящая статья рассматривает такие сравнительно-статистические методы идентификации как «проверка на вхождение» и «пересечение множеств».

Текст при таких методах анализируется как последовательность отдельных составляющих его элементов (символов, словоформ, грамматических классов и т. д.) или групп элементов длиной N–N-грамм. Кроме того, исследованию подвергаются такие лексико-квантитативные проявления как биграммы и триграммы символов, наиболее частые слова языков.

Перед тем, как выполнять эти задачи, необходимо проделать процедуры предобработки текста:

- Нормализация и приведение слов к стандартному виду, принятому в языках типологической группы
- Обработку текста документа при помощи программы-морфоанализатора
- Автоопределение языка посредством специальной программы
- Определение тональности документа (входит в разряд второстепенных задач).
- Распознавание именных сущностей, содержащихся в документе (носит факультативный характер)

«Проверка на вхождение» заключается в проверке наличия в тексте заранее маркированных слов, предполагающих принадлежность текста к искомой тематике. Процедуру проверки на вхождение можно выразить в виде алгоритма:

%Пусть [’a’,’b’,’c’] — это пространство признаков текста. Пусть ’a’ — признак, маркирующий какое-либо свойство текста (например, содержание элементов из эталонного списка). Тогда на языке программного кода:

```
if 'a' in ['a','b','c']:
    print «В анализируемом тексте содержатся элементы из множества А. Возможно, текст носит искомый характер. «
    else:
    print «В исследуемом тексте не обнаружено элементов из эталонного списка»
```

При обнаружении вхождения эталонных ключевых слов необходимо переходить к следующему этапу: проверке через работу со множествами. Множество в Python – это “контейнер”, содержащий не повторяющиеся элементы идущие в случайном порядке [4]. С множествами можно выполнять множество операций: находить объединение, пересечение, разность и симметрическую разность.

Перечислим ряд операций по сравнению множеств «mnozh1» и «mnozh2» в языке программирования Python:

% len(mnozh1) :число элементов в множестве mnozh1(размер множества).

% x in mnozh1 :принадлежит ли x множеству mnozh1

% mnozh1.isdisjoint(mnozh2) :истина, если mnozh1 и mnozh2 не имеют общих элементов

% mnozh1 == mnozh2 :все элементы mnozh1 принадлежат mnozh2, все элементы mnozh2 принадлежат mnozh1.

% mnozh1.issubset(mnozh2) или mnozh1<= mnozh2 :все элементы mnozh1 принадлежат mnozh2.

% mnozh1.issuperset(mnozh2) или mnozh1 >= mnozh2 :аналогично

% mnozh1.union(mnozh2, ...) mnozh1 | mnozh2 | ... :объединение нескольких множеств.

% mnozh1.intersection(mnozh2, ...) или mnozh1 & mnozh2 & ... :пересечение mnozh1 и mnozh2.

% mnozh1.difference(mnozh2, ...) mnozh1 - mnozh2 : множество из всех элементов set, не принадлежащие ни одному из other.

% mnozh1.symmetric_difference(mnozh2) или mnozh1 ^ mnozh2 :множество из элементов, встречающихся в одном множестве, но не встречающиеся в обоих.

Прикладное значение для идентификации текстов будет иметь выявление пересечения множеств ключевых слов текста и заранее разработанного "эталонного" множества ключевых слов.

Как правило с точки зрения структурно-статистического подхода тексты религиозного содержания содержат в себе большой процент сигнальных слов. Эти сигнальные слова необходимо оформить в так называемый эталонный список. Наличие в каком-либо тексте таких сигнальных слов позволяет выдвинуть предположение, что текст относится к религиозной, а в некоторых случаях и экстремистской тематике.

Источником слов, включаемых в «эталонный список», являются «прецедентные тексты». Ю.Н. Караулов определяет прецедентные тексты следующим образом: «Назовем прецедентными – тексты, (1) значимые для той или иной личности в познавательном и эмоциональном отношениях, (2) имеющие сверхличностный характер, т. е. хорошо известные и широкому окружению данной личности, включая ее предшественников и современников, и, наконец, такие, (3) обращение к которым возобновляется неоднократно в дискурсе данной личности» [5]. Содержание в анализируемом тексте отрывков из прецедентных текстов поможет идентифицировать текст [6].

Прецедентные тексты в цифровом формате хранятся в составе «лингвистических ресурсов» (Linguistic Data) [7]. Лингвистические ресурсы (ЛР) – это множество определенным образом организованных речевых и языковых данных, находящихся на машинных носителях информации и используемых в различных сферах практической деятельности (образовании, промышленности, экономике, культуре, искусстве, издательстве). В самом общем виде ЛР – это своеобразные лингвистические базы данных [8, 9], которые можно обновлять и в которых можно искать ту или иную информацию. Лингвистические ресурсы необходимы как пользователям ПК, так и различным компьютерным системам, связанным с обработкой текста речи: реферирование, аннотирование и перевод текстов, автоматический анализ текста, синтез речи и текста.

Таким образом, для того чтобы успешно выполнять операции проверки на вхождение и операции со множествами, необходимо вначале набрать корпуса прецедентных текстов, а затем выделить их них ключевую терминологию, с помощью которой можно маркировать и анализируемые тексты.

Филологи-русисты уже определили прецедентные тексты религиозного содержания для выделения ключевых высказываний для христианских текстов [10, 11]. Подобную работу по источниковедению прецедентных текстов ислама предстоит проделать и востоковедам-исламоведам.

Собранные востоковедами корпуса прецедентных текстов на исламскую тематику в дальнейшем должны быть обработаны на предмет выделеня ключевых «сигнальных» слов. Списки выделенных ключевых сигнальных слов в дальнейшем должны быть оформлены в виде базы данных, которые и будут источником для сравнительно-статистических методов «проверка на вхождение» и «пересечение множеств».

Таким образом, задачами востоковедов-исламоведов для использования структурно-статистических методов в ходе мониторинга и идентификации текстов являются:

- Сбор корпусов прецедентных текстов по исламской тематике на различных языках [12].
- Выявление из корпусов прецедентных текстов по исламской тематике сигнальных слов, наличие которых в тексте позволяет экспертам сделать выводы о включении текста в определенную категорию. Формирование «эталонного множества» сигнальных слов для выполнения операций «проверка на вхождение» и «поиск пересечения множеств» [13].
- Составление базы данных сигнальных слов по исламской тематике.
- Разработка алгоритмов и скриптов для проведения операций «Проверка на вхождение» и «Поиск пересечения множеств».

Литература

1. Жукова, О.С. Информационный экстремизм как угроза безопасности Российской Федерации / О.С. Жукова // Вестник Воронежского института МВД России. – 2007. – Т. 1.

2. Кокорев, В.Г. Понятие и признаки религиозного экстремизма / В.Г. Кокорев // Социально-экономические явления и процессы. – 2014. – Т. 5.

3. Шибаев, М.В. Манипулятивное использование прецедентных текстов в религиозном дискурсе / М.В. Шибаев // Вестник Красноярского государственного педагогического университета им. В.П. Астафьева. – 2013. – Т. 3.

4. Прохоренок, Н. Python 3. Самое необходимое / Н. Прохоренок. – БХВ-Петербург, 2016.

5. Караулов, Ю.Н. Русский язык и языковая личность / Ю.Н. Караулов, Д.Н. Шмелев. – М.: Наука, 1987.

6. Бобырева, Е.В. Прецедентные высказывания религиозного дискурса / Е.В. Бобырева // Известия Волгоградского государственного педагогического университета. – 2007. – Т. 2.

7. Chiarcos, C. Towards Open Data for Linguistics: Linguistic Linked Data / C. Chiarcos // New Trends of Research in Ontologies and Lexical Resources. – 2013. – P. 7–25.

8. Мишанкина, Н.А. Базы данных в лингвистических исследованиях / Н.А. Мишанкина // Вопросы лексикографии. – 2013. – Т. 1 (3).

9. Мишанкина, Н.А. Технология баз данных в социогуманитарных исследованиях / Н.А. Мишанкина // Гуманитарная информатика. – 2012. – Т. 6.

10. Мишланов, В.А. Диалогичность церковно-религиозных текстов / В.А. Мишланов, В.А. Салимовский // Вестн. Перм. ун-та. – 2010. – Т. 6, № 12. – С. 24–28.

11. Мишланов, В.А. Этнический экстремизм в массовой коммуникации с точки зрения проблем судебной лингвистической экспертизы / В.А. Мишланов, В.А. Салимовский // Вестн. Перм. ун-та. – 2013. – Т. 4, № 24. – С. 63–75.

12. Saad, M. OSAC: Open Source Arabic Corpora / M. Saad, W. Ashour // 6th International Conference on Electrical and Computer Systems (EECS'10), Nov 25-26, 2010, Lefke, Cyprus. 2010. – P. 118–123.

13. Фаткулин, Б.Г. Использование лингвистически ориентированных модулей на языке python для обработки больших текстовых массивов на восточных языках в целях эффективного сбора и обработки данных по отраслям востоковедческой тематики (на примере NLTK) / Б.Г. Фаткулин // Вестник ЮУрГУ. Серия «Лингвистика». – 2015. – Т. 12, № 1. – С. 72–75.

Фаткулин Булат Гилимдарович, кандидат филологических наук, доцент кафедры общей лингвистики, Южно-Уральский государственный университет (г. Челябинск), fatkulinbg@susu.ru

Поступила в редакцию 25 июня 2016 г.

DOI: 10.14529/ling160304

USE OF SET THEORY IN COMPARATIVE LOGICAL METHODS OF SELECTING TEXTS ON ISLAMIC CONTENT IN NETWORK RESOURCES

***B.G. Fatkulin**, fatkulinbg@susu.ru*
South Ural State University, Chelyabinsk, Russian Federation

The procedure of identifying texts on Islamic content in network resources can be carried out using methods of comparative logic «check on entry» and «intersection of sets.» For any language the preassembled corpus of precedent texts on Islamic content is to be joint. The list of key concepts is extracted from these corpora with the help of statistical methods. The list of the selected keyword appears in the form of a set. The entity of key concepts can be represented as a database and can be used as a reference set A. The texts under expertise will be presented as a set of concepts B. The set A is compared with the set B. The variety of the set B elements from the set A, the nature and degree of the intersection of the two sets allows us to identify the text as having an Islamic character. Python programming language provides great opportunities for working with strings, tuples, dictionaries, and sets. In the process of monitoring network resources and searching for texts on Islamic content the methods outlined in this article, as well as a database with the reference lists of keywords can be used.

Keywords: information extremism, forensic linguistic examination, network resources, monitoring network resources, comparative statistical methods, Python language, sets, linguistic resources, precedent texts, Islam.

References

1. Zhukova O.S., Ivanchenko R.B., Trukhachev V.V. Informacionnyj ehkstremizm kak ugroza bezopasnosti Rossijskoj Federacii [Information Extremism as a Threat to the Security of the Russian Federation]. *Bulletin of the Voronezh Institute of Russian Ministry of Internal Affairs.* 2007. No. 1.

2. Kokorev V.G.. Ponyatie i priznaki religioznogo ehkstremizma [The Concept and Signs of the Religious Extremism]. *Socio-economic Phenomena and Processes.* 2014. No. 5.

3. Shibayev M.V. Manipulyativnoe ispol'zovanie precedentnyh tekstov v religioznom diskurse [Manipulative Use of Precedent Texts in Religious Discourse]. *Bulletin of the Krasnoyarsk State Pedagogical University.* 2013. No. 3.

4. Prohorenok N. Python 3. Samoe neobhodimoe [Python 3. The Most Important Things]. *BHV-Petersburg*, 2016.

5. Karaulov J.N., Shmelev D.N. *Russkij yazyk i yazykovaya lichnost'* [Russian Language and Linguistic Identity]. Moscow, Nauka, 1987.

6. Bobyreva E.V. Precedentnye vyskazyvaniya religioznogo diskursa [Precedent Sayings in Religious Discourse]. *Proceedings of the Volgograd State Pedagogical University*. 2007. No. 2.

7. Chiarcos C. and etc. Towards Open Data for Linguistics: Linguistic Linked Data. *New Trends of Research in Ontologies and Lexical Resources.* 2013. pp. 7–25

8. Mishankina N.A. Bazy dannyh v lingvisticheskih issledovaniyah [Databases in Linguistic research]. *Problems of Lexicography.* 2013. № (3).

9. Mishankina N.A. Tekhnologiya baz dannyh v sociogumanitarnyh issledovaniyah [Database Technology in Social and Humanitarian Studies]. *Humanitarian Informatics.* 2012. No. 6.

10. Mishlanov V.A., Salimovsky V.A. Dialogichnost' cerkovno-religioznyh tekstov [Dialogs in Church and Religious Texts]. *The Bulletin of Perm University.* 2010, no. 6, pp. 24–28.

11. Mishlanov V.A., Salimovsky V.A. Ethnic Extremism in Mass Communication from the Perspective of Forensic Linguistic Examination of Problems. *Vestnik. Perm. Univ.* 2013, no. 24, pp. 63–75.

12. Saad M., Ashour W. OSAC: Open Source Arabic Corpora. *6th International Conference on Electrical and Computer Systems (EECS'10), Nov 25–26, 2010,* Lefke, Cyprus, 2010, pp. 118–123.

13. Fatkulin B.G. Ispol'zovanie lingvisticheski orientirovannyh modulej na yazyke python dlya obrabotki bol'shih tekstovyh massivov na vostochnyh yazykah v celyah ehffektivnogo sbora i obrabotki dannyh po otraslyam vostokovedcheskoj tematiki (na primere NLTK) [The Use of Linguistically-oriented Modules in Python to Handle Large Arrays of Text in Oriental Languages for the Purpose of Efficient Collection and Oriental Languages Data Processing (using NLTK as an example)]. *Bulletin of South Ural State University. Ser. Linguistics.* 2015, v. 12, no. 1, pp 72–75.

Received 25 June 2016

Fatkulin Bulat, Candidate Degree in Philology, associate professor, Chair of General Linguistics, South Ural State University (Chelyabinsk), fatkulinbg@susu.ru

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Фаткулин, Б.Г. Использование теории множеств в сравнительно-логических методах выделения текстов на исламскую тематику в процессе мониторинга сетевых ресурсов / Б.Г. Фаткулин // Вестник ЮУрГУ. Серия «Лингвистика». – 2016. – Т. 13, № 3. – С. 22–26. DOI: 10.14529/ling160304

FOR CITATION

Fatkulin B.G. Use of Set Theory in Comparative Logical Methods of Selecting Texts on Islamic Content in Network Resources. *Bulletin of the South Ural State University. Ser. Linguistics*. 2016, vol. 13, no. 3, pp. 22–26. (in Russ.). DOI: 10.14529/ling160304